Henryk Banaszak*

# Skalowanie kumulatywne – zarys problematyki

## Idea skalowania kumulatywnego

### Model deterministyczny

#### Skala dystansu społecznego Bogardusa

Początki skalowania kumulatywnego sięgają przełomu lat trzydziestych ubiegłego stulecia i wiążą się z osobą Emory'ego Bogardusa (1926, 1928, 1933), badającego uprzedzenia etniczne społeczeństwa amerykańskiego za pomocą skonstruowanej przez siebie *skali dystansów społecznych*. Wybierał on grupę odniesienia (etniczną, religijną, zawodową) i pytał respondentów, jak *blisko siebie* skłonni są dopuścić członków tej grupy. Jako wskaźników społecznej bliskości (społecznego dystansu) używał odpowiedzi na pytanie o akceptację obecności „obcego" w swoim otoczeniu. W diagnozie poziomu dystansu społecznego względem grupy odniesienia Bogardus używał zazwyczaj kilku pytań, na które można było odpowiedzieć *tak* lub *nie*:

| lp | Czy akceptuje Pan(i) [*członka grupy odniesienia*] w roli: | Dystans społeczny względem grupy |
|---|---|---|
| 1 | Spokrewnionego z Panem(ią) w wyniku małżeństwa | 1 |
| 2 | Pana(i) bliskiego przyjaciela | 2 |
| 3 | Pana(i) sąsiada mieszkającego na tej samej ulicy | 3 |
| 4 | Osoby wykonującej ten sam zawód co Pan(i) | 4 |
| 5 | Obywatela Pana(i) kraju | 5 |
| 6 | Turysty odwiedzającego Pana(i) kraj | 6 |

* Henryk Banaszak ukończył socjologię na Uniwersytecie Warszawskim, gdzie się doktoryzował. W Instytucie Socjologii UW prowadzi zajęcia ze statystyki. Niezależny konsultant statystyczny, ekspert MEN i CKE (habanasz@is.uw.edu.pl).

Zgodnie z „geometryczną" interpretacją dystansu społecznego, osoba, która dopuszcza „obcego" *blisko siebie* (na przykład w roli męża lub bliskiego krewnego) nie powinna mieć nic przeciwko jego obecności *daleko od siebie* (na przykład w roli obywatela lub turysty).

Jeśli zatem *kolejność* dystansów społecznych w kwestionariuszu jest dobrana właściwie, odpowiedzi respondentów powinny układać się w charakterystyczny wzór, stanowiący esencję „kumulatywności":

- każdy, kto akceptuje „obcego" w *pewnym* oddaleniu od siebie, akceptuje jego obecność w każdym oddaleniu *większym.*

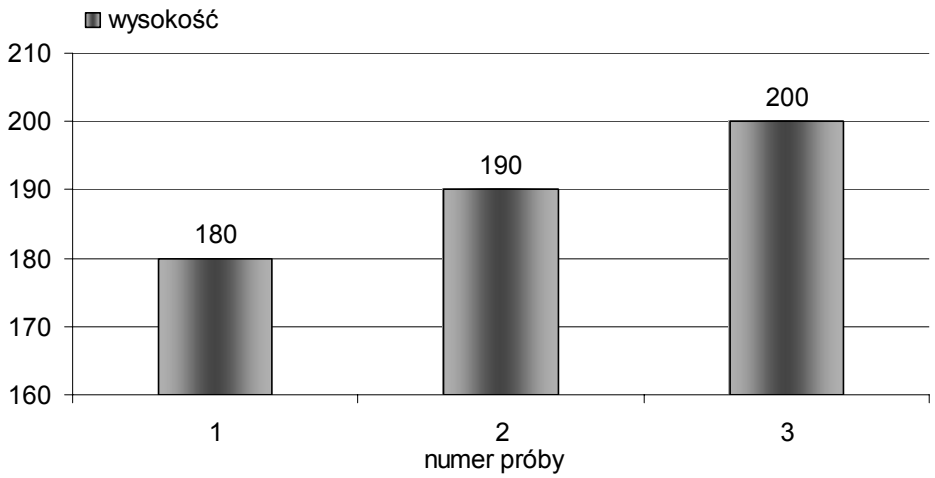
Gdyby wszyscy respondenci zachowywali się zgodnie z tym wzorem, znajomość liczby odpowiedzi „tak" udzielanych na powyższe pytania pozwoliłaby dla konkretnego odpowiadającego dokładnie określić, jak odpowiadał na każde z pytań, a zarazem od którego poziomu bliskości zaczął odpowiadać „tak", zatem, jaki jest jego dystans względem grupy odniesienia. Numer porządkowy takiego pytania byłby wówczas wartością cechy ukrytej diagnozowanej przez powyższy zestaw pytań.

Efektywność tak skonstruowanej skali zależy oczywiście od tego, do jakiego stopnia spełnione są założenia dotyczące uporządkowania wskaźników (pytań) i sposobu ich traktowania przez respondentów. Warto więc wiedzieć, co dokładnie założenia te oznaczają, czyli jakie są logiczne i ontologiczne podstawy skalowania kumulatywnego. Formalizację tych założeń umożliwiającą rzeczową nad nimi dyskusję zawdzięczamy Leo Guttmanowi. W niniejszym artykule przedstawimy własną wersję tej formalizacji i wprowadzimy model Guttmana w kategoriach probabilistycznych, co pozwoli lepiej zrozumieć jego specyfikę i podobieństwa z innymi modelami skalowania kumulatywnego. Formalizację modeli poprzedzimy analizą prostego przykładu.

## Skok wzwyż

Dobrym sposobem pokazania formalnych własności deterministycznej wersji modelu skalowania kumulatywnego jest przykład nieco uproszczonego konkursu skoku wzwyż.

Załóżmy, że w konkursie są tylko 3 wysokości, a każdy uczestnik ma tylko jedną próbę na każdej z nich. Ponadto, do próby na wysokości następnej dopuszczani są tylko ci, którzy pokonali wysokość poprzednią.




**Rysunek 1.** Wysokości w skoku wzwyż jako próby w teście skoczności

Potraktujmy próby na kolejnych wysokościach jak pytania testu diagnozującego „skoczność" zawodników. Na każde z nich można odpowiedzieć poprawnie (przeskoczyć wysokość) lub niepoprawnie (nie przeskoczyć jej). Ponadto, z istoty konkursu wynika, że próby są *naturalnie uporządkowane* ze względu na stopień ich trudności. Z przyjętych reguł wynika ponadto, że nie można pokonać wysokości *wyższej* nie pokonując uprzednio wysokości *niższej* – zawodnik, który nie przeskoczył którejś z wysokości odpada z konkursu, a jego próby na wysokościach następnych uznaje się za nieudane.

Jeśli oznaczyć przez *1* udaną próbę przeskoczenia wysokości, a przez *0* próbę nieudaną, to okaże się, że każdy z uczestników konkursu, w zależności od tego, jakie wyniki w poszczególnych próbach uzyskał, może należeć tylko do jednej z czterech kategorii:

Z tych samych założeń wynika także, że do ustalenia wyniku uczestnika konkursu nie jest potrzebna znajomość jego wyników na poszczególnych wysokościach – wystarczy znać *liczbę udanych prób*, gdyż ta jednoznacznie określa numer ostatniej wysokości, którą skoczek pokonał. Zauważmy, że ta możliwość diagnozowania „skoczności" zawodnika stosuje się również do sytuacji, gdy nie znamy *dokładnie* wysokości, które próbują przeskoczyć

**Tabela 1.** Dopuszczalne wyniki uproszczonego konkursu skoku wzwyż

| Kategoria zawodnika | Próba 1 (180 cm) | Próba 2 (190) | Próba 3 (200) | Liczba udanych prób |
|---|---|---|---|---|
| a | 0 | 0 | 0 | 0 |
| b | 1 | 0 | 0 | 1 |
| c | 1 | 1 | 0 | 2 |
| d | 1 | 1 | 1 | 3 |

uczestnicy konkursu, a tylko mamy pewność, że są *uporządkowane*, że każda następna jest wyższa od poprzedniej. W konsekwencji zliczając ich udane próby *porządkujemy* uczestników ze względu na poziom ich skoczności, nie precyzując jakiej wysokości odpowiada każda z pozycji tego porządku.

## Filary kumulatywnego modelu skalowania cechy ukrytej

Opisany wyżej konkurs dobrze ilustruje schemat skalowania kumulatywnego. Próby na kolejnych wysokościach to obserwowalne wskaźniki cechy ukrytej, którą jest „skoczność" zawodnika. Liczba udanych prób pozwala zawodników uporządkować ze względu na poziom ich cechy ukrytej bez względu na to, na jakich wysokościach te próby przeprowadzano.

Możliwość tak prostego uporządkowania uczestników konkursu ze względu na ich poziom cechy ukrytej opiera się na trzech założeniach, stanowiących filary kumulatywnego skalowania cechy ukrytej:

**Założenie o współmierności poziomów cechy ukrytej i poziomów trudności prób**

(i) sukces lub porażka w każdej z prób zależy wyłącznie od poziomu cechy ukrytej próbującego: próba kończy się sukcesem, gdy jej poziom trudności jest niższy niż poziom cechy ukrytej próbującego.

**Założenie o uporządkowaniu prób testowych**

(ii) kolejne próby są uporządkowane ze względu na stopień ich trudności dla uczestników – sukces w każdej następnej próbie wymaga od uczestnika wyższego poziomu cechy ukrytej niż poziom wystarczający do sukcesu w próbie poprzedniej.

**Założenie kumulatywności reakcji**

(iii) uczestnicy dopuszczani są do kolejnych prób tylko wtedy, gdy wszystkie poprzednie próby zakończyły się sukcesem.

Zauważmy, że sformułowane wyżej reguły mają charakter bezwyjątkowy: każdy uczestnik reaguje zgodnie z założeniem (i), każda próba spełnia postulat (ii) i każdy uczestnik przestrzega założenia (iii). Założenia zawierają *de* facto duże kwantyfikatory, stąd opisywany przez nie model nazywamy deterministycznym – poziom cechy ukrytej uczestnika determinuje jednoznacznie rezultat każdej z jego prób, a w konsekwencji liczbę prób udanych, czyli wartość poziomu jego cechy ukrytej.

## Model probabilistyczny

### Skok wzwyż w wersji probabilistycznej

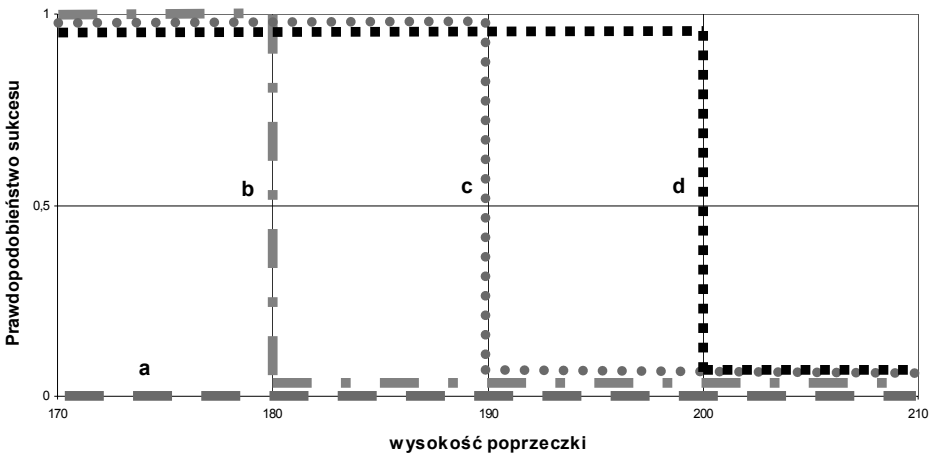
Formalizację własności modeli skalowania poprzedzimy wprowadzeniem kluczowego dla modeli probabilistycznych pojęcia *funkcji reakcji* na wskaźnik.

Funkcja reakcji[1] dla każdego poziomu cechy ukrytej i dla każdego binarnego pytania-wskaźnika określa prawdopodobieństwo udzielenia na to pytanie jednej z dwóch możliwych odpowiedzi, zazwyczaj nazywanej „odpowiedzią poprawną" lub „sukcesem". Odpowiedź poprawną wygodnie jest oznaczać liczbą *1*, a wtedy liczba *0* reprezentuje „odpowiedź błędną" lub „porażkę".

Zobaczmy, jak za pomocą funkcji reakcji daje się opisać nasz konkurs skoku wzwyż, w którym przecież mamy do czynienia z sukcesami i porażkami. Najpierw przedstawimy w kategoriach probabilistycznych charakterystykę czterech kategorii uczestników konkursu, ***a***, ***b***, ***c*** i ***d*** .

Skoczność zawodnika typu ***a*** jest poniżej poziomu trudności pierwszego progu testu – nie jest on w stanie przeskoczyć wysokości 180 (choć być może byłby w stanie pokonać wysokość 175) i w związku z tym nie jest dopuszczony do dalszych prób. Prawdopodobieństwo pokonania wysokości wyższych niż pierwsza wynosi dla niego 0.

[1] W literaturze anglojęzycznej występuje jako IRF – *Item Response Function* lub ICC – *Item Characteristic Curve.*




**Rysunek 2.** Probabilistyczna charakterystyka skoczności czterech typów zawodników

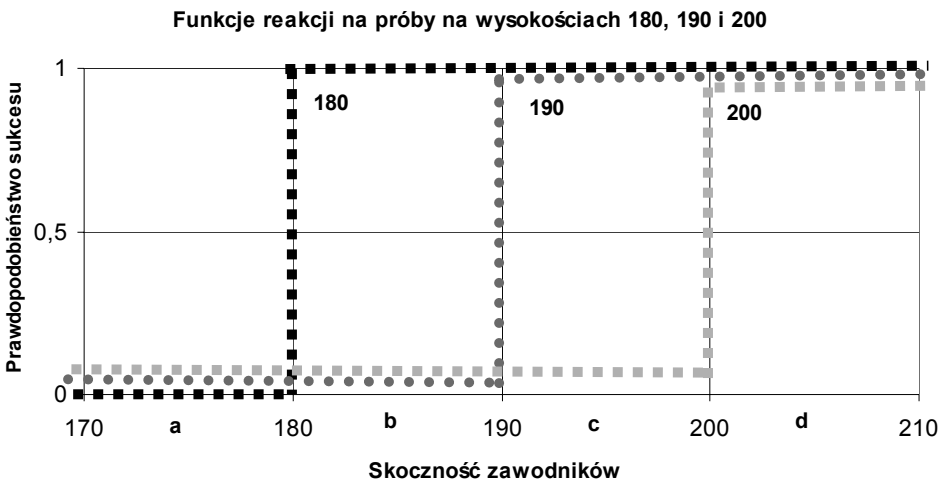
Zawodnik typu ***b*** ma poziom skoczności wystarczający do pokonania wysokości 180, być może także 185, lecz jego skoczność nie pozwala na pokonanie wysokości 190 i następnych. Zatem prawdopodobieństwo pomyślnego zakończenia pierwszej próby (180) wynosi dla niego 1, natomiast prawdopodobieństwa sukcesu w próbach następnych są już zerowe.

Analogicznie, prawdopodobieństwo pokonania pierwszych dwóch prób, 180 i 190, dla zawodnika typu ***c*** wynosi 1, lecz prawdopodobieństwo przeskoczenia wysokości 200 jest równe 0.

Zwycięzcami konkursu są zawodnicy typu ***d***, którzy z prawdopodobieństwem 1 pokonują każdą z wysokości. Być może byliby oni w stanie pokonać wysokości powyżej 200, jednak konkurs kończy się właśnie na tej próbie.

Zwróćmy uwagę na dwa szczegóły powyższego opisu. Po pierwsze, pozwala on rozróżnić tylko cztery ***typy*** zawodników, nie przesądza natomiast, *ilu* uczestników konkursu zostanie zaliczonych do każdego z nich. Bez względu na to ilu uczestników „znajdzie się" w każdym z typów, wszystkie osoby tego samego typu z punktu widzenia tego testu będą traktowane jako osoby posiadające tę samą skoczność. Po drugie, widać, jakie są ograniczenia testu przeprowadzanego w powyższy sposób – nie jest on w stanie diagnozować poziomów skoczności pozwalających przeskakiwać wysokości niższe od 180 i wyższe od 200, a ponadto jest to test bardzo zgrubny – pozwala rozróżniać poziomy skoczności z dokładnością nie większą niż 10.

Prawdopodobieństwa pokonania poszczególnych wysokości przez zawodników każdego z typów dają się równoważnie przedstawić za pomocą *funkcji reakcji*, jeśli zmienimy interpretację osi poziomej wykresu. Zgodnie z zasadą współmierności teraz będzie ona reprezentowała poziom skoczności uczestnika konkursu. Wartości odkładane na osi pionowej są znów prawdopodobieństwami pokonania wysokości przez zawodnika o poziomie skoczności odłożonym na osi poziomej. Funkcja reakcji (prawdopodobieństwa przeskoczenia wysokości) określona będzie dla trzech prób, różniących się stopniem trudności. Funkcja reakcji na próbę pokazuje zatem prawdopodobieństwo sukcesu dla wszystkich potencjalnych uczestników konkursu, którzy mogą mieć poziom skoczności odłożony na osi poziomej.




**Rysunek 3.** Funkcje reakcji zawodników skaczących wzwyż na próby na trzech wysokościach

Prawdopodobieństwo sukcesu na wysokości 180 dla zawodnika typu ***a***, którego skoczność jest niższa niż 180 wynosi 0. Zarazem nie wiadomo, jaka jest rzeczywista skoczność zawodnika typu ***a*** – konkurs daje nam prawo do stwierdzenia iż leży gdzieś (nie wiadomo gdzie) na lewo od punktu 180.

Prawdopodobieństwo sukcesu w pierwszej próbie osiąga wartość 1 dla zawodnika typu ***b***, którego skoczność jest równa przynajmniej 180. Zarazem, zawodnik tego typu ma zerowe prawdopodobieństwa osiągnięcia sukcesu w obu próbach na wysokościach 190 i 200.

Skoczność zawodnika typu ***c*** leży między 190 a 200, zatem z prawdopodobieństwem 1 pokonuje on próby na wysokościach 180 i 190, lecz ponieważ

jego skoczność jest niższa niż 200, prawdopodobieństwo jej pokonania wynosi dla niego 0. Zawodnik typu ***d*** na pewno przeskakuje wszystkie trzy wysokości, lecz ponieważ konkurs do nich się ogranicza, nie wiadomo, jaka jest granica jego możliwości – na podstawie tego konkursu można ustalić tylko, że wynosi ona co najmniej 200.

Zauważmy, że powyższy wykres nie zmieni wyglądu, gdy na osi poziomej zamiast skoczności wyrażonej w wysokościach (180, 190, 200) użyjemy numerów prób: 1, 2, 3, uporządkowanych od najłatwiejszej do najtrudniejszej. Jeśli nie znamy wysokości poszczególnych prób, nie możemy *ilościowo* diagnozować skoczności uczestników testu, wciąż jednak możemy ich ze względu na tę skoczność uporządkować, ustalając dla każdego z nich liczbę prób zakończonych sukcesem.

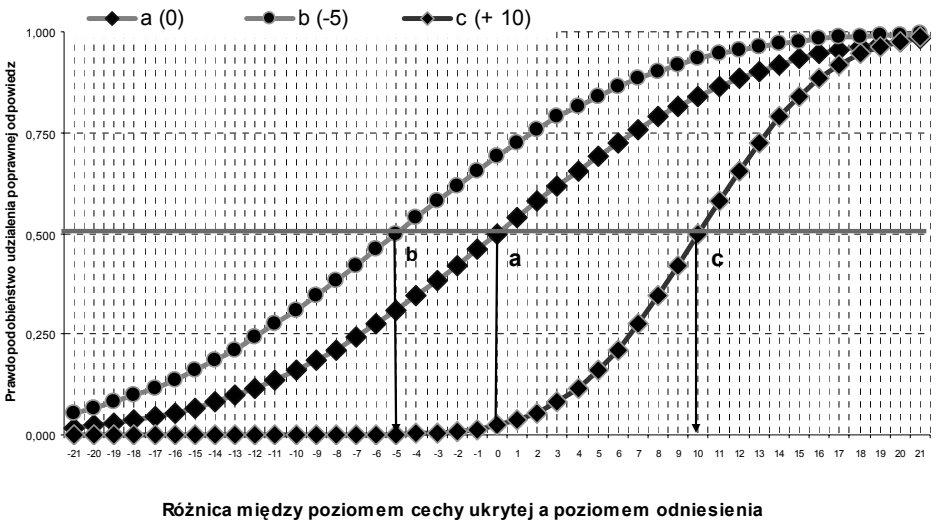
Powyższy wykres, mimo iż posługuje się pojęciem prawdopodobieństwa, przedstawia funkcje reakcji dla deterministycznego modelu skalowania kumulatywnego, nazywanego skalogramem Guttmana, który szczegółowo omawiamy dalej. Determinizm tego modelu polega na tym, że dla każdego uczestnika testu i dla każdej z prób prawdopodobieństwa sukcesu wynoszą albo 1 albo 0. Wobec tego ***każda osoba***, która sukcesem zakończyła próbę o trudności *k*, odniosła również sukces w ***każdej z prób*** o trudności niższej niż *k*.

Uchylenie dużego kwantyfikatora z powyższego zdania prowadzi do probabilistycznej wersji modelu skalowania kumulatywnego. Uchylenie to polega, mówiąc obrazowo, na „wyginaniu" funkcji reakcji i ulokowaniu punktu ich przecięcia z poziomem 0,5 nad tym punktem osi trudności, który charakteryzuje poziom cechy ukrytej uczestnika testu.

## Funkcja reakcji w probabilistycznej wersji skalowania kumulatywnego

Posłużmy się przykładem testu składającego się z trzech pytań: ***a***, ***b*** i ***c***. Poziom cechy ukrytej wyrażony jest jako odchylenie od pewnego poziomu odniesienia.

„Wygięte" funkcje reakcji są rosnącymi funkcjami poziomu cechy ukrytej. Są one skonstruowane tak, aby prawdopodobieństwo poprawnej odpowiedzi na pytanie reprezentowane przez funkcję wynosiło 0,5 dla tych osób, których poziom umiejętności jest równy poziomowi trudności pytania. Pytania testu dają się zatem uporządkować ze względu na to, w którym miejscu




**Rysunek 4.** Funkcje reakcji dla pytań *b*, *a*, *c* różniących się poziomem trudności

nad osią trudności ich funkcja reakcji przecina poziom 0,5, a zatem ze względu na poziom ich trudności. W naszym przykładzie najłatwiejsze jest pytanie ***b***, potem ***a***, najtrudniejsze jest pytanie ***c*** . Porządek ten jest jednoznacznie określony także wtedy, gdy wielkości odkładane na osi poziomej zostaną monotonicznie przekształcone.

Nieprzypadkowo funkcje reakcji dla trzech pytań w powyższym przykładzie są różne: mają różny kształt (nachylenie) i dla różnych argumentów przyjmują wartość 0,5. Różne warianty probabilistycznych modeli skalowania różnią się bowiem sposobem definiowania funkcji reakcji i relacji, w jakich pozostają względem siebie funkcje reprezentujące poszczególne wskaźniki.

To co wspólne wszystkim klasycznym modelom skalowania kumulatywnego[2], powyższy wykres pokazuje:

- im wyższy przyjąć poziom cechy ukrytej, tym, dla każdego pytania, prawdopodobieństwo udzielenia poprawnej odpowiedzi rośnie.
- dla dowolnego poziomu cechy ukrytej porządek pytań ustalony ze względu na wielkość prawdopodobieństwa udzielenia na nie poprawnych odpowiedzi jest taki sam.
- porządek pytań jest zgodny z porządkiem poziomów cechy ukrytej, dla których prawdopodobieństwo poprawnej odpowiedzi wynosi 0,5.

[2] Z nielicznymi wyjątkami – patrz Uebersax (1999, 23: 283–297).

### Lokalna niezależność reakcji

Dotychczasowe rozważania zajmowały się tymi założeniami modelu, który dotyczył relacji między wartościami cechy ukrytej i rozpatrywanych osobno wartościami wskaźników. Zajmijmy się teraz pokrótce własnościami *łącznego rozkładu* wszystkich zmiennych definiujących model.

Wspólną cechą probabilistycznych wersji modeli skalowania kumulatywnego[3] jest założenie *lokalnej niezależności reakcji*. W skrócie daje się ono sformułować tak: proces reagowania na wskaźniki jest procesem bez pamięci. Oznacza to w szczególności, że:

- poziom cechy ukrytej osoby reagującej na wskaźniki jest taki sam bez względu na ich kolejność „podawania",
- prawdopodobieństwa „poprawnych" reakcji na kolejne wskaźniki zależą wyłącznie od odległości między poziomem cechy ukrytej respondenta i poziomem „trudności" wskaźników,
- prawdopodobieństwo serii reakcji na wskaźniki dla pojedynczej osoby jest równe iloczynowi prawdopodobieństw reakcji na każdy ze wskaźników z osobna.

Pojedyncza osoba ma swoją wartość cechy ukrytej. Z kolei każdy wskaźnik ma swoją funkcję reakcji. Osoba reaguje na wskaźnik z prawdopodobieństwem wyznaczanym przez jego funkcję. Po udzieleniu odpowiedzi, bez względu na to, czy była „poprawna" czy nie, osoba „zapomina przeszłość" i reaguje na kolejny wskaźnik z prawdopodobieństwem zależnym wyłącznie od jego krzywej reakcji. Odpowiada to wykonywaniu przez osobę tylu rzutów monetami, ile wskaźników zawiera test, przy czym prawdopodobieństwa wyrzucenie „orła" („sukcesu", czy inaczej „poprawnej reakcji") są za każdym razem różne i wyznaczane przez funkcję reakcji wskaźnika odpowiadającego kolejnemu rzutowi.

## Probabilistyczne modele skalowania kumulatywnego – podsumowanie

Probabilistyczne modele skalowania kumulatywnego przyjmują założenia, które adekwatnie, jak sądzimy, opisują sytuację diagnozowania poziomu

[3] Jak wyżej.

umiejętności (kompetencji) za pomocą zestawu zadań testowych. Ich naturalnym terenem zastosowań są zatem szkolne testy kompetencyjne. Zestawienie założeń tych modeli i ich interpretację dla sytuacji testowania osiągnięć przedstawia poniższa tabela.

**Tabela 2.** Założenia probabilistycznych modeli skalowania kumulatywnego a sytuacja testowania kompetencji

| Założenia probabilistycznego modelu skalowania kumulatywnego | Przykład: osoby rozwiązujące zadania testowe |
|---|---|
| Obiekty różnią się parametrami istotnymi dla wyniku zdarzenia losowego | Osoby różnią się poziomem kompetencji, **łatwością**, z jaką rozwiązują zdania testowe |
| Wskaźniki różnią się parametrami istotnymi dla wyniku zdarzenia losowego | Pytania testowe różnią się stopniem **trudności**, jaką sprawiają odpowiadającym, |
| Obserwowalna reakcja obiektu na wskaźnik jest zdarzeniem losowym | Osoby testowane reagują do pewnego stopnia **przypadkowo**: osoba bardzo kompetentna może nie odpowiedzieć na pytanie łatwe, a osoba mało kompetentna może odpowiedzieć na pytanie trudne |
| Zbiór możliwych reakcji i ich prawdopodobieństwa stanowią zmienną losową, której rozkład zależy od parametrów osoby i parametrów wskaźnika (zmienna losowa – funkcja reakcji) | Szanse na poprawną odpowiedź osoby na pojedyncze pytanie testowe **zależą zarazem** od tego, jak trudne jest to pytanie i jak kompetentna jest odpowiadająca na nie osoba |
| Reakcje obiektów o ustalonych parametrach (tym samym poziomie kompetencji) są stochastycznie niezależne | Pojedyncza osoba odpowiada na kolejne pytanie testu „**bez pamięci**" o wynikach poprzednich odpowiedzi i wyłącznie w zależności od tego, jak trudne jest kolejne pytanie i jak kompetentna jest osoba |

# Model skalowania w zapisie formalnym

W niniejszym artykule ograniczymy się do modeli skalowania kumulatywnego, w których wskaźniki są dychotomiczne (binarne). Modelem skalowania dla wskaźników dychotomicznych nazywać będziemy piątkę:

$$\langle \Omega, X, \boldsymbol{\beta}, \delta, P_{\beta\delta} \rangle, \tag{1}$$

w której:

- $\Omega = \{\omega_1, \omega_2, \omega_3, ..., \omega_v, ..., \omega_n\}$ jest *n*-elementowym zbiorem obiektów,
- X jest *k*-elementowym zbiorem binarnych wskaźników $(X_1, X_2, X_3, ..., X_i, ..., X_k)$,
- **β** jest jednowymiarową zmienną ukrytą określoną w Ω,
- **δ** jest c*k*-elementowym wektorem parametrów wskaźników $(X_1, ..., X_k)$, gdzie c=1, 2, 3, ... oznacza liczbę parametrów pojedynczej funkcji reakcji; **δ** można też traktować jako funkcję rzeczywistą, która wskaźnikom przyporządkowuje ich parametry o wartościach rzeczywistych,
- $P_{\beta\delta}$ jest funkcją reakcji wiążącą prawdopodobieństwo $P(X_{iv}=x)$, $x \in \{0,1\}$ reakcji obiektu $\omega_v$ na wskaźnik $X_i$ z poziomem cechy ukrytej obiektu $\beta(\omega_v)$ oraz poziomem trudności wskaźnika $\delta(X_i)$.

W dalszym ciągu tekstu przyjmiemy „szkolną" semantykę modelu skalowania, stąd $\Omega=\{\omega_1, \omega_2, ..., \omega_n\}$ nazywany będzie zbiorem *osób* odpowiadających na *pytania* wskaźnikowe $(X_1, X_2, X_3, ..., X_i, ..., X_k)$, na które można udzielić odpowiedzi „*poprawnej*" $(X_{iv}=1)$ lub „*niepoprawnej*" $(X_{iv}=0)$, zaś prawdopodobieństwo „sukcesu" lub „porażki" zależy od poziomu „*umiejętności osoby*" $\beta(\omega_v)$ oraz od „*trudności pytania*" $\delta(X_i)$.

## Parametry osób i wskaźników

W modelach, w których funkcja reakcji na wskaźnik ma jawną postać, jest ona funkcją parametrów osób i parametrów wskaźników. Oba parametry, $\beta(\omega_v)$ i $\delta(X_i)$ zdają sprawę z położenia wskaźników i osób na wspólnej skali.

W oznaczeniach stosowanych dalej indeksy *v* oraz *i* porządkują poziomy cechy ukrytej (*umiejętności*) oraz parametry wskaźnika (*trudności*) zgodnie ze swymi wartościami. Zatem jeśli $a > b$, to $\beta(\omega_a) > \beta(\omega_b)$, a jeśli c > d, to $\delta(X_c) > \delta(X_d)$.

## Funkcja reakcji: kumulatywność i lokalna niezależność

W zależności od modelu, funkcja reakcji może mieć postać jawną lub niejawną. Jak się przekonamy, w różnych modelach nakładane są na nią różne założenia, choć zawsze prawdopodobieństwo udzielenia poprawnej odpowie-

dzi na pytanie $X_i$ przez osobę $\omega_v$ , oznaczane przez $P(X_{vi}=1)$, ma taką postać, że jest ono niemalejącą funkcją zmiennej **β**, a reakcje na poszczególne wskaźniki w grupach osób o tym samym poziomie umiejętności są od siebie stochastycznie niezależne (lokalna niezależność reakcji)

$$P(X_{vi} = x_{vi} \mid \beta_v) = \prod_{i=1}^{k} P(X_{vi} = x_{vi} \mid \beta_v) \tag{2}$$

Zasada kumulatywności wyrażona w wersji probabilistycznej oznacza, że reakcje na wskaźniki są stochastycznie pozytywnie zależne. Ma ona wówczas postać nierówności:

$$\delta_j > \delta_i \Rightarrow \quad \forall v\, P(X_{vi} = 1 \mid X_{vj} = 1) \geq P(X_{vi} = 1) \tag{3}$$

# Skalogram Guttmana

## Założenia

Mimo swej deterministycznej natury model Guttmana daje się przedstawić jako model probabilistyczny, w którym prawdopodobieństwa są równe 0 lub 1. Model ten określony jest przez trzy założenia:

- (*porządek osób*) Osoby różnią się pod względem poziomu „umiejętności" (**β**) i można je ze względu na tę cechę uporządkować.
- (*porządek wskaźników*) Wskaźniki różnią się ze względu na stopień „trudności" (**δ**) i można je ze względu na tę własność uporządkować.
- (*kumulatywność reakcji*) **Każdy**, kto zareagował pozytywnie/poprawnie na wskaźnik o pewnym stopniu trudności, reaguje pozytywnie/poprawnie na **wszystkie** łatwiejsze wskaźniki:

$$\forall i, j(\delta_j > \delta_i) \Rightarrow \forall v, P(X_{vi} = 1 \mid X_{jv} = 1) = 1 \tag{4}$$

## Konsekwencje

### Porządek w zbiorze osób

Z powyższych założeń wynika (co ilustrował przykład konkursu skoku wzwyż), że **numer najtrudniejszego pytania**, na które osoba odpowiedziała pozytywnie, porządkuje osoby w ten sam sposób co liczba poprawnych odpowiedzi udzielonych przez osoby i w ten sam sposób co poziom cechy ukry-

tej. Zatem używając nieważonej sumy binarnych zmiennych wskaźnikowych możemy wyznaczyć porządek osób ze względu na poziom wartości posiadanej przez nie cechy ukrytej (poziom umiejętności).

## Porządek w zbiorze wskaźników

Konsekwencją zasady kumulatywności jest możliwość uporządkowania wskaźników ze względu na stopień ich „trudności". Porządek ten wyznacza liczba osób, które pozytywnie odpowiedziały na dany wskaźnik, czyli suma wartości zmiennej wskaźnikowej,

## Ograniczenie klasy dopuszczalnych profili reakcji

**Profilem reakcji** nazywać będziemy wektor składający się z zer i jedynek, o długości równej liczbie wskaźników, reprezentujący reakcje osoby na każdy z nich. Przy *k* wskaźnikach liczba wszystkich możliwych profili wynosi $2^k$. Z założeń modelu Guttmana wynika, że **nie wszystkie profile reakcji są dopuszczalne**, a liczba profili dopuszczalnych, dla których spełniona jest zasada kumulatywności reakcji, wynosi $k+1$.

Konsekwencję tę obrazuje tabela 3, w której przedstawione są wszystkie możliwe profile odpowiedzi dla trzech wskaźników. Profile należące do klasy profili dopuszczalnych oznaczono kolorem szarym.

**Tabela 3.** Możliwe profile reakcji dla przypadku trzech wskaźników

| | **Wskaźnik** | | | **Liczba pozytywnych odpowiedzi** |
|---|---|---|---|---|
| Profil | $X_1$ | $X_2$ | $X_3$ | $\sum_{i=1}^{3} X_i$ |
| A | 1 | 1 | 1 | 3 |
| B | 1 | 1 | 0 | 2 |
| C | 1 | 0 | 1 | 2 |
| D | 0 | 1 | 1 | 2 |
| E | 1 | 0 | 0 | 1 |
| F | 0 | 1 | 0 | 1 |
| G | 0 | 0 | 1 | 1 |
| H | 0 | 0 | 0 | 0 |

Profile dopuszczalne ($P_D$) oznaczone są kolorem szarym
Profile niedopuszczalne ($P_N$) oznaczone są kolorem białym

Kategoria profili reakcji zgodnych z założeniem kumulatywności reakcji odgrywa istotną rolę we wszystkich modelach skalowania kumulatywnego. Jak się dalej okaże, profile te mają największe prawdopodobieństwo zaistnienia w klasach ukrytych, wyznaczanych przez statystykę dostateczną w modelu Rascha.

### Strukturalne zero

Kolejną konsekwencją kumulatywności reakcji w wersji deterministycznej jest **występowanie strukturalnego zera w łącznych rozkładach par wskaźników.** Jeżeli profile reakcji należą do klasy dopuszczalnych, w zbiorze osób nie ma takich, które odpowiadają na wskaźnik trudniejszy nie odpowiedziawszy na wskaźnik łatwiejszy. Oznacza to, że jeśli wskaźnik $X_i$ jest łatwiejszy od wskaźnika $X_j$ , w ich łącznym rozkładzie empirycznym liczebność w komórce II powinna być równa 0 (patrz tabela 4).

**Tabela 4.** Rozkład łączny wskaźników $(X_i, X_j)$

| | | $X_{j\ trudny}$ | | Suma |
|---|---|---|---|---|
| | | 0 | 1 | |
| $X_{i\ łatwy}$ | 0 | I | II | 1-p |
| | 1 | III | IV | p |
| Suma | | 1-q | q | 1,00 |

Jak widać, oba wskaźniki są wtedy stochastycznie zależne.

## Własności modelu Guttmana

### Przykład skalowania

Deterministyczny charakter modelu Guttmana jest źródłem jego nieusuwalnych wad. Pokażemy to na prostym przykładzie procesu wyznaczania łącznego rozkładu wskaźników i cechy ukrytej w teście złożonym z trzech pytań wskaźnikowych $X_1$, $X_2$, $X_3$. Ich łączny rozkład w próbie (empiryczny rozkład profili reakcji) umieszczony jest na dole tabeli 5.

**Tabela 5.** Przykładowy model Guttmana dla skalowania cechy ukrytej za pomocą trzech dychotomicznych wskaźników

| | | Model | | | | | | Oczekiwane częstości profili | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | $\Sigma_j P(X_1=x_1, X_2=x_2, X_3=x_3 \mid \boldsymbol{\beta}=\beta_j) P(\boldsymbol{\beta}=\beta_j)$ | | | | | | | |
| Rozkład **β** | | Dopuszczalne profile | | | $P(X_i=1 \mid \boldsymbol{\beta}=\beta_j)$ | | | 0,30 | 0,35 | 0,30 | 0,05 | 0,00 | 0,00 | 0,00 | 0,00 |
| $\beta_j$ | $P(\boldsymbol{\beta}=\beta_j)$ | $X_1$ | $X_2$ | $X_3$ | $X_1$ | $X_2$ | $X_3$ | **111** | **110** | **100** | **000** | **101** | **011** | **010** | **001** |
| 3 | 0,30 | 1 | 1 | 1 | 1,00 | 1,00 | 1,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| 2 | 0,35 | 1 | 1 | 0 | 1,00 | 1,00 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| 1 | 0,30 | 1 | 0 | 0 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| 0 | 0,05 | 0 | 0 | 0 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | | Empiryczne częstości brzegowe wskaźników | | | ← próba → | | | Empiryczny rozkład profili reakcji | | | | | | | |
| | | $P(X_1=1)$ | $P(X_2=1)$ | $P(X_3=1)$ | | | | $p_{111}$ | $p_{110}$ | $p_{100}$ | $P_{000}$ | $p_{101}$ | $p_{011}$ | $p_{010}$ | $p_{001}$ |
| | | 0,70 | 0,65 | 0,55 | ← | | | 0,30 | 0,20 | 0,10 | 0,05 | 0,10 | 0,05 | 0,10 | 0,10 |
| | | | | | | | | Stopień zgodności empirycznego rozkładu profili z częstościami oczekiwanymi | | | | | | | |
| | | Różnica między empirycznymi częstościami profili w próbie a częstościami oczekiwanymi przez model | | | | | | *0,00* | *–0,15* | *–0,20* | *0,00* | *0,10* | *0,05* | *0,10* | *0,10* |

Najpierw z rozkładu wyznacza się $P(X_i=1)$, empiryczne brzegowe częstości wartości 1 dla każdego ze wskaźników, co *wyznacza ich porządek* ze względu na poziom ich „trudności". Przypomnijmy, z założeń modelu Guttmana wynika, że porządek wskaźników ze względu na ich trudność jest zgodny z ich uporządkowaniem ze względu na częstość poprawnych odpowiedzi: im większa część próby odpowiada poprawnie na pytanie, tym jest ono „łatwiejsze".

Sumując liczbę poprawnych odpowiedzi w profilu reakcji wyznacza się (szacuje) brzegowe częstości $P(\boldsymbol{\beta}=\beta_j)$, składające się na *brzegowy rozkład cechy ukrytej* $\boldsymbol{\beta}$. Zauważmy, że w szacunku tym używamy wszystkich profili reakcji, także tych, które z punktu widzenia założeń modelu nie powinny wystąpić w rozkładzie empirycznym[4].

Następnie, z założenia lokalnej niezależności reakcji wyznacza się warunkowe częstości poszczególnych profili w każdej z klas cech ukrytej:

$$P(X_1=x_1, X_2=x_2, X_3=x_3|\boldsymbol{\beta}=\beta_j) = P(X_1=x_1|\boldsymbol{\beta}=\beta_j)\, P(X_2=x_2|\boldsymbol{\beta}=\beta_j)\, P(X_3=x_3|\boldsymbol{\beta}=\beta_j) \qquad (5)$$

W modelu Guttmana warunkowe prawdopodobieństwa wskaźników definiujące ich funkcje reakcji są równe albo 0 albo 1, wobec tego ich iloczyny również mogą przyjmować te dwie wartości, co widać w prawej górnej części tabeli. Niezerowe są warunkowe prawdopodobieństwa tylko tych profili, które są zgodne z profilem dopuszczalnym dla danej klasy ukrytej.

Z warunkowych prawdopodobieństw profili reakcji (od **111** do **001**) dla każdego poziomu $\beta_j$ cechy ukrytej $\boldsymbol{\beta}$ oraz z jej brzegowych częstości $P(\boldsymbol{\beta}=\beta_j)$ można wyznaczyć cały łączny rozkład cechy ukrytej i obserwowalnych wskaźników, czyli wszystkie częstości łączne postaci:

$$P(X_1=x_1, X_2=x_2, X_3=x_3, \boldsymbol{\beta}=\beta_j) = P(X_1=x_1, X_2=x_2, X_3=x_3|\boldsymbol{\beta}=\beta_j)\, P(\boldsymbol{\beta}=\beta_j) \qquad (6)$$

Sumując je w kolumnach odpowiadających poszczególnym profilom uzyskujemy częstości profili reakcji oczekiwane przy założeniach modelu Guttmana.

---

[4] Posłużyliśmy się mechaniczną regułą zliczania poprawnych odpowiedzi ignorując tym samym wątpliwości, czy operacja ta jest uzasadniona dla profili niedopuszczalnych. Odmiany modelu Guttmana różnią się między innymi sposobem rozwiązywania problemu przyporządkowania do klasy ukrytej profili niedopuszczalnych.

Porównanie rozkładu profili reakcji oczekiwanego wedle modelu z rozkładem empirycznym wykazuje, że tylko 35% badanych reagowało na wskaźniki zgodnie z zasadą kumulatywności reakcji, częstości oczekiwane są równe zarejestrowanym tylko dla profili **111** oraz **000**. Tylko dla tej grupy badanych poziom umiejętności jest wyznaczony jednoznacznie, dla pozostałych 65% uczestników testu decyzja o poziomie cechy ukrytej musi być rozstrzygana arbitralnie.

## Podsumowanie

Sprawdźmy teraz, jak model Guttmana radzi sobie z rozwiązywaniem zasadniczych problemów skalowania.

### I. Problem skalowalności

Porównanie empirycznych i oczekiwanych częstości profili umożliwia ocenę stopnia zgodności łącznego rozkładu wskaźników z założeniami modelu. Jak widać, w ostatnim dolnym wierszu tabeli 5 empiryczne częstości profili reakcji odbiegają od oczekiwanych, miejscami znacznie. W modelu Guttmana nie ma jednak kryteriów pozwalających zdecydować, czy rozbieżności te są zbyt duże czy też nie. Rozbieżności te wyraża się przy użyciu kategorii *błędu*, czyli reakcji niezgodnej z zasadą kumulatywności. Błędy reakcji zlicza się przy tym na dwa sposoby: licząc *błędne profile* (z klasy niedopuszczalnych) albo zliczając *błędy w profilach*. W obu wariantach otrzymuje się współczynniki skalowalności, których wartości oddzielające sytuacje „skalowalności" od pozostałych przyjmuje się arbitralnie[5].

### II. Problem liczby wymiarów cechy ukrytej i relacji między nimi

W modelu Guttmana jest to problem nierozwiązywalny z założenia. Zdarza się, co prawda, że niska wartość współczynnika skalowalności skłania badacza do poszukiwania wyjaśnień tego faktu, lecz odbywają się one bez naruszania aksjomatyki modelu. Poszukiwania takie zmierzają zazwyczaj do

[5] Warianty skalogramu Guttmana różnią się między innymi sposobem wyrażania stopnia zgodności danych z modelem. Gdy dodać do siebie empiryczne częstości profili niedopuszczalnych, otrzymany odsetek badanych, którzy nie spełniają założenia kumulatywności reakcji. Stopień zgodności można także liczyć liczbą inwersji niezbędnych do uzyskania profilu dopuszczalnego z niedopuszczalnego.

sprawdzenia, czy porządek wskaźników określany przez ich brzegowe rozkłady jest taki sam we wszystkich podzbiorach zbioru obiektów. Wystarczy, na przykład, aby kobiety inaczej postrzegały „dystanse społeczne" lub „naganność zachowania" niż mężczyźni, aby porządek „trudności" pytań wskaźnikowych wyznaczany dla całej zbiorowości był inny niż w każdym z jej podzbiorów. W rezultacie, mimo iż w obu grupach osoby reagują zgodnie z zasadą kumulatywności, w badaniu rejestruje się wiele profili niedopuszczalnych, gdyż porządek wskaźników wyznaczany dla całej zbiorowości jest inny niż w każdym z jej podzbiorów.

**III. Czy wszystkie wskaźniki są potrzebne?**

W modelu Guttmana nie funkcjonuje akceptowane kryterium pozwalające oddzielić wskaźniki potrzebne od zbędnych. Operacje typu „co by było, gdyby ten wskaźnik usunąć", które rutynowo wykonuje się w analizie danych empirycznych, w tym wypadku prowadzą do sprawdzania, czy współczynnik skalowalności po takim zabiegu poprawił się czy też nie. W obu sytuacjach decyzje o losie wskaźnika są jednak podejmowane „na wyczucie".

**IV. Jakie są własności diagnostyczne poszczególnych wskaźników?**

We wszystkich modelach skalowania kumulatywnego na podstawie binarnych wskaźników występuje parametr nazywany „poziomem trudności" pytania wskaźnikowego[6]. Uniwersalny postulat prostoty, a nie konstrukcja modelu Guttmana sugerują, aby test nie zawierał wskaźników o identycznym poziomie trudności. Innych parametrów własności diagnostycznych pytań wskaźnikowych skalogram Guttmana nie przewiduje.

**V. Jak skalować – funkcja agregująca profile**

W modelu Guttmana wskaźnikiem poziomu cechy ukrytej jest liczba „poprawnych" odpowiedzi na pytania wskaźnikowe. Taka funkcja skalująca wyznacza właściwe poziomy cechy ukrytej dla profili z klasy dopuszczalnych przez model, zawodzi jednak dla profili niedopuszczalnych: tę samą liczbę poprawnych odpowiedzi można odnotować odpowiadając tylko na pytania najłatwiejsze albo tylko na najtrudniejsze. Dla osoby, której profil reak-

[6] Jeśli interpretować „poziom trudności" pytania jako wartość oczekiwaną odsetka maksymalnej punktacji, to stosuje się ją również do wskaźników politomicznych.

cji należy do klasy niedopuszczalnych liczba poprawnych odpowiedzi nie jest zatem wiarygodnym wskaźnikiem poziomu ukrytej cechy a wtedy przyporządkowanie profilowi poziomu cechy musi być arbitralne. Liczba „poprawnych" odpowiedzi jest zatem dobrą funkcją skalującą tylko dla osób reagujących zgodnie z zasadą kumulatywności.

W problemie stopnia skalowalności i jednoznaczności wyznaczania poziomu cechy ukrytej, deterministyczna wersja modelu nie pozwala podejmować decyzji uzasadnianych inaczej niż za pomocą „reguły kciuka", mimo prób nadania rozwiązaniom pozorów racjonalności (por. Stookey, Baer 1976 lub McIver, Carmines 1981). Racjonalne, czyli dobrze uzasadnione, decyzje w sprawie agregacji profilu w poziom cechy ukrytej możliwe są po przeformułowaniu deterministycznej zasady kumulatywności Guttmana w jej probabilistyczny odpowiednik, w którym profile z klasy niedopuszczalnych stają się zjawiskiem dopuszczanym przez założenia modelu. Nieparametryczną probabilizację skalogramu Guttmana zaproponował holenderski statystyk R.J. Mokken (1997).

# Skalogram Mokkena

## Notacja

Z założenia współmierności cechy ukrytej, własności osoby $\omega_v$, oraz poziomu trudności, parametru wskaźnika $X_i$ wynika, że oba rodzaje obiektów, osoby i pytania, są reprezentowane przez punkty tej samej osi liczb rzeczywistych $\mathfrak{R}$. Oznaczmy zgodnie z tym założeniem atrybuty osób i wskaźników:

$$\begin{aligned} \beta(\omega_v) &= \beta_v,\ \beta_v \in \mathfrak{R} \\ \delta(X_i) &= \delta_i,\ \delta_i \in \mathfrak{R} \end{aligned} \tag{7}$$

Zatem $\beta_v$ oznacza poziom cechy ukrytej, „umiejętności" osoby $\omega_v$, zaś $\delta_i$ oznacza parametr wskaźnika $X_i$ interpretowany jako poziom jego „trudności".

## Założenia

Model Mokkena w wersji „klasycznej", nazywanej modelem z podwójną monotonicznością, ma cztery założenia:

## Punkt równości atrybutów osób i wskaźników

Osoba $\omega_v$ może udzielić pozytywnej ($X_{iv}=1$) lub negatywnej ($X_{iv}=0$) odpowiedzi na wskaźnik $X_i$. W modelu Mokkena funkcja reakcji wiążąca prawdopodobieństwo $P(X_{iv}=1)$ z wartościami $\beta_v$ i $\delta_i$ zdefiniowana jest następująco:

- jeżeli $\beta_v < \delta_i$ to, $0 \leq P(X_{iv}=1) < 0{,}5$
- jeżeli $\beta_v = \delta_i$ to $0 < P(X_{iv}=1) < 0{,}5$,
- jeżeli $\beta_v > \delta_i$ to, $0{,}5 < P(X_{iv}=1) \leq 1$

Funkcja reakcji przybiera zatem wartość 0,5, gdy $\beta_v = \delta_i$. Oznacza to, że osoba $\omega_v$, której poziom umiejętności $\beta_v$ jest taki sam jak poziom trudności $\delta_i$ pytania $X_i$ ma takie same szanse odpowiedzieć na nie „poprawnie" jak i „niepoprawnie".

## Monotoniczność względem cechy ukrytej

Prawdopodobieństwo pozytywnej reakcji na dowolny wskaźnik $X_i$ nie maleje wraz ze wzrostem pozycji osoby na skali ukrytej ($\beta_v$). Osoba o pewnym poziomie umiejętności ma wyższe prawdopodobieństwo udzielenia poprawnej odpowiedzi to samo pytanie, niż osoba, której poziom umiejętności jest niższy. Ponadto, jest tak dla wszystkich pytań. Innymi słowy, funkcja reakcji osoby na wskaźnik ze względu na pozycję osoby na skali, przy stałej trudności wskaźnika jest niemalejąca względem **β**. Oznacza to, że

$$\forall i,l,m \;\; \beta_l \leq \beta_m \Rightarrow P(X_{li}=1) \leq P(X_{mi}=1) \tag{8}$$

Zatem, jeżeli wszystkie wskaźniki reprezentują tę samą zmienną ukrytą, wtedy uporządkowanie osób za pomocą prawdopodobieństwa pozytywnej reakcji powinno być takie samo dla wszystkich wskaźników (Schuur 2003: 145).

## Monotoniczność względem trudności wskaźnika

Każda osoba ma wyższe prawdopodobieństwo udzielenia poprawnej odpowiedzi na pytanie łatwiejsze niż na pytanie trudniejsze. Oznacza to, że funkcja reakcji na wskaźnik ze względu na trudność wskaźnika przy stałym poziomie cechy tej pozycji osoby jest niemalejąca[7] względem **δ**.

[7] Przyjmuje się także silniejsze założenia o ścisłej monotoniczności tych funkcji.

$$\forall v,i,j \;\; \delta_i \le \delta_j \Rightarrow P(X_{vi}=1) \ge P(X_{vj}=1) \tag{9}$$

### Lokalna niezależność reakcji

Niech **X** = **x** oznacza wektor reprezentujący oznaczane liczbą 0 lub 1 odpowiedzi na pytania wskaźnikowe ($X_1$, $X_2$ ...., $X_k$). Wektor taki nazywać będziemy profilem odpowiedzi. Wektor odpowiedzi osoby $\omega_v$ oznaczymy przez $\mathbf{X}_v = \mathbf{x}_v$

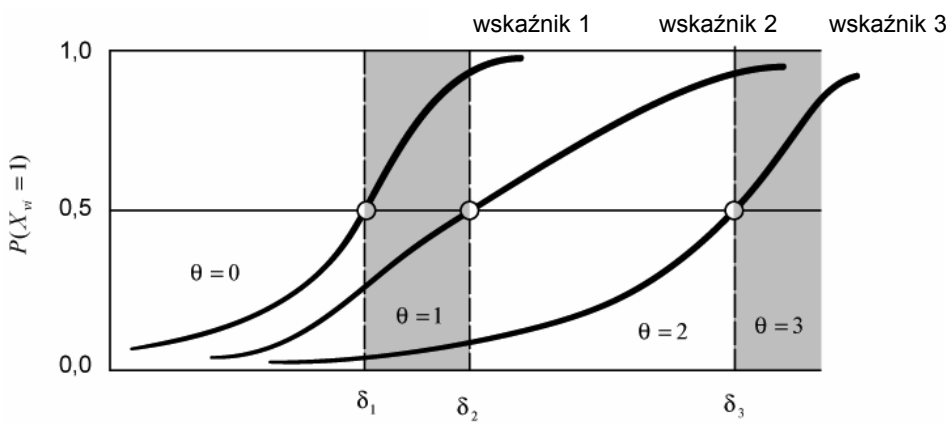
Reakcje na wskaźniki osób o tym samym poziomie cechy ukrytej $\beta_v$ są stochastycznie niezależne.

$$P(\mathbf{X}_v = \mathbf{x}_v \mid \beta_v) = \prod_{i=1}^{k} P(X_{vi} = x_{vi} \mid \beta_v) \tag{10}$$

## Konsekwencje

### Funkcje reakcji

Krzywe reakcji poszczególnych wskaźników dają się uporządkować ze względu na punkt na osi poziomej, nad którym przecinają poziom 0,5, jak na rysunku 5[8].




**Rysunek 5.** Krzywe reakcji na trzy wskaźniki dla skalogramu Mokkena

Na rysunku tym stopień trudności wskaźników określają stałe $\delta_1$, $\delta_2$ i $\delta_3$. Funkcje reakcji dzielą zbiorowość testowanych na cztery klasy[9], definiowane

[8] Rysunek jest autorstwa Jakuba Komorka (2006).

[9] Podobnie jak na rysunku 4.

przez punkty, nad którymi prawdopodobieństwo poprawnej odpowiedzi na wskaźnik przekracza 0,5. Do klasy {β=0} należą badani, których prawdopodobieństwo poprawnej odpowiedzi wynosi mniej niż 0,5 dla każdego ze wskaźników, do klasy {β=1} należą ci, dla których to prawdopodobieństwo przekracza 0,5 tylko dla wskaźnika 1. Ich poziom cechy ukrytej leży między $\delta_1$ i $\delta_2$. I tak dalej, analogicznie jak na rysunku 4.

Z założenia podwójnej monotoniczności wynika, że krzywe reakcji wskaźników w skalogramie Mokkena nie przecinają się. Ponadto, podobnie jak w modelu Guttmana porządek wskaźników (poziom ich trudności) jest wyznaczany przez brzegowe częstości poprawnych na nie odpowiedzi i jest on taki sam w każdej z klas wartości cechy ukrytej. Oznacza to, że warunkowe prawdopodobieństwa modelu Mokkena spełniają charakterystyczny wzór, który zilustrujemy przedstawiając skalowanie cechy ukrytej metodą Mokkena w tej samej co poprzednio zbiorowości, w której model Guttmana tak dramatycznie zawiódł.

Porządek wskaźników oraz rozkład cechy ukrytej są takie same jak poprzednio, gdyż zostały wyznaczone w ten sam sposób[10]. Warunkowe prawdopodobieństwa reakcji na wskaźniki tworzące podwójnie monotoniczne funkcje reakcji układają się w charakterystyczny wzór:

- (dla każdego wskaźnika) im niższy jest poziom cechy ukrytej, tym warunkowe prawdopodobieństwo poprawnej reakcji na wskaźnik maleje (w kolumnach prawdopodobieństwa maleją od góry w dół),
- (dla każdego poziomu cechy ukrytej) im trudniejszy jest wskaźnik tym niższe jest prawdopodobieństwo udzielenia nań poprawnej odpowiedzi (prawdopodobieństwa maleją w wierszach od lewej do prawej, w miarę jak rośnie trudność wskaźnika).

Wartości funkcji reakcji spełniające oba postulaty monotoniczności wpisaliśmy dla przykładu starając się jak najwyraźniej pokazać wzór, w który się układają. Mimo to, dzięki probabilizacji modelu oczekiwane częstości profili są znacznie bliższe częstościom empirycznym. W łącznym rozkładzie obserwowalnych wskaźników i cechy ukrytej nie ma już komórek z zerowymi prawdopodobieństwami, gdyż model dopuszcza istnienie profili z klasy niedopuszczalnych Guttmana. Dzięki temu różnica częstości profili oczekiwanych wedle modelu i częstości empirycznych reakcji wedle tego modelu zna-

[10] O uzasadnieniu sposobu wyznaczania częstości cechy ukrytej piszemy dalej.

**Tabela 6.** Przykładowy model skalowania cechy ukrytej za pomocą trzech dychotomicznych wskaźników w probabilistycznej wersji Mokkena

| | | Model | | | | | | Oczekiwane częstości profili | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Funkcje reakcji | | | $\Sigma_j P(X_1=x_1, X_2=x_2, X_3=x_3 \mid \boldsymbol{\beta}=\beta_j) P(\boldsymbol{\beta}=\beta_j)$ | | | | | | | |
| Rozkład **β** | | Profile Guttmana | | | $P(X_i=1 \mid \boldsymbol{\beta}=\beta_j)$ | | | 0,33 | 0,21 | 0,10 | 0,04 | 0,13 | 0,08 | 0,06 | 0,04 |
| $\beta_j$ | $P(\boldsymbol{\beta}=\beta_j)$ | $X_1$ | $X_2$ | $X_3$ | $X_1$ | $X_2$ | $X_3$ | 111 | 110 | 100 | 000 | 101 | 011 | 010 | 001 |
| 3 | 0,30 | 1 | 1 | 1 | 0,90 | 0,80 | 0,70 | 0,50 | 0,22 | 0,05 | 0,01 | 0,13 | 0,06 | 0,02 | 0,01 |
| 2 | 0,35 | 1 | 1 | 0 | 0,80 | 0,70 | 0,60 | 0,34 | 0,22 | 0,10 | 0,02 | 0,14 | 0,08 | 0,06 | 0,04 |
| 1 | 0,30 | 1 | 0 | 0 | 0,70 | 0,60 | 0,50 | 0,21 | 0,21 | 0,14 | 0,06 | 0,14 | 0,09 | 0,09 | 0,06 |
| 0 | 0,05 | 0 | 0 | 0 | 0,50 | 0,40 | 0,30 | 0,06 | 0,14 | 0,21 | 0,21 | 0,09 | 0,06 | 0,14 | 0,09 |
| | | Empiryczne częstości brzegowe wskaźników | | | ← próba → | | | Empiryczny rozkład profili reakcji | | | | | | | |
| | | $P(X_1=1)$ | $P(X_2=1)$ | $P(X_3=1)$ | | | | $p_{111}$ | $p_{110}$ | $p_{100}$ | $P_{000}$ | $p_{101}$ | $p_{011}$ | $p_{010}$ | $p_{001}$ |
| | | 0,70 | 0,65 | 0,55 | ← | | | 0,30 | 0,20 | 0,10 | 0,05 | 0,10 | 0,05 | 0,10 | 0,10 |
| | | | | | | | | Stopień zgodności empirycznego rozkładu profili z częstościami oczekiwanymi | | | | | | | |
| | | Różnica między empirycznymi częstościami profili w próbie a częstościami oczekiwanymi przez model | | | | | | *–0,03* | *–0,01* | *0,00* | *0,01* | *–0,03* | *–0,03* | *0,04* | *0,06* |

cząco zmalała – gdyby zsumować moduły takich różnic, okazałoby się, że suma ta wynosi tylko 0,19 w porównaniu z wartością 0,65 charakteryzującą model Guttmana.

## Zależność wskaźników

Ze względu na probabilistyczny charakter modelu, w łącznych rozkładach par wskaźników nie musi – jak w modelu Guttmana – występować strukturalne zero. Z założeń **(i)-(iv)** wynika natomiast, że wskaźniki są parami pozytywnie skorelowane (Mokken 1971). Dla wskaźników z przykładu z tabeli 6 konsekwencja ta nie jest spełniona:

**Tabela 7.** Zależności między wskaźnikami a oczekiwania modelu Mokkena

| $X_1\backslash X_2$ | 0 | 1 | |
|---|---|---|---|
| 0 | 0,15 | 0,15 | |
| 1 | 0,20 | 0,50 | 0,70 |
| | | 0,65 | |

| $X_1\backslash X_3$ | 0 | 1 | |
|---|---|---|---|
| 0 | 0,15 | 0,15 | |
| 1 | 0,30 | 0,40 | 0,70 |
| | | 0,55 | |

| $X_2\backslash X_3$ | 0 | 1 | |
|---|---|---|---|
| 0 | 0,15 | 0,20 | |
| 1 | 0,30 | 0,35 | 0,65 |
| | | 0,55 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| $\rho_{12}$ | 0,21 | | $\rho_{13}$ | 0,07 | $\rho_{23}$ | -0,03 |

W żadnym z rozkładów pary wskaźników nie ma strukturalnego zera, oczekiwanego w modelu Guttmana, pamiętajmy jednak, że warunek ten wynikał z deterministycznej zasady kumulatywności reakcji. Nasz przykładowy rozkład jest jednak sprzeczny z jej wersją probabilistyczną – wbrew oczekiwaniom wynikającym z założenia podwójnej monotoniczności modelu Mokkena wskaźniki $X_2$ oraz $X_3$ nie są pozytywnie skorelowane, Odsetek osób, które udzieliły poprawnej odpowiedzi na $X_3$ nie odpowiadając zarazem poprawnie na łatwiejsze pytanie $X_2$, jest zbyt wysoki i kowariancja obu zmiennych jest ujemna.

Kolejną konsekwencją założenia podwójnej monotoniczności jest charakterystyczny wzór relacji między komórkami łącznych rozkładów par wskaźników. Wprowadźmy oznaczenie pozwalające ten wzór opisać. Niech $\mathbf{P}^{00}$ oznacza symetryczną kwadratową macierz o wymiarze $k \times k$, której elementy $\mathbf{P}^{00}(i,j)$ zawierają empiryczne częstości $P(X_i=0, X_j=0)$ udzielenia obu „niepoprawnych" odpowiedzi na pytania wskaźnikowe $X_i$ oraz $X_j$. Analogicznie macierz $\mathbf{P}^{11}$ zawiera empiryczne częstości $P(X_i=1, X_j=1)$ udzielenia obu od-

powiedzi „poprawnych". Jeśli reakcje osób testowanych spełniają dokładnie założenia podwójnej monotoniczności reakcji, to elementy obu macierzy układają się we wzór podobny do tablicy z funkcjami reakcji: po uporządkowaniu wskaźników ze względu na trudność (od najłatwiejszych do najtrudniejszych) wartości macierzy $\mathbf{P}^{00}(i,j)$ wzrastają (nie maleją) w miarę wzrostu $j$ a wartości $\mathbf{P}^{11}(i,j)$ maleją (nie rosną). Dla wskaźników z przykładu z tabeli 6 macierze te mają następującą postać:

**Tabela 8.** Macierze $P^{11}$ oraz $P^{00}$ dla danych z tabeli 6

| | | | $\mathbf{P}^{00}$ | |
|---|---|---|---|---|
| | | $X_1$ | $X_2$ | $X_3$ |
| | $X_1$ | | 0,15 | 0,15 |
| $\mathbf{P}^{11}$ | $X_2$ | 0,50 | | 0,15 |
| | $X_3$ | 0,40 | 0,35 | |

Jak widać, relacje między fragmentami łącznych rozkładów wskaźników $X_1$, $X_2$, $X_3$ z naszego przykładu spełniają – choć w słabszej wersji – oczekiwania wynikające z założeń modelu Mokkena.

## Statystyka dostateczna cechy ukrytej β

Model Mokkena jest probabilistyczny, lecz jednocześnie nieparametryczny, gdyż postulaty nakładane na funkcje reakcji mają nieparametryczny, porządkowy charakter. Model pozwala zatem porządkować osoby ze względu na poziom cechy ukrytej, do którego zaklasyfikuje je funkcja agregująca profile, nie pozwala natomiast interpretować odległości między poziomami tej cechy. Pozwala natomiast interpretować pozycje osób zaklasyfikowanych przez funkcję agregującą do jakiejś klasy, jako odpowiadające poziomowi cechy ukrytej leżącemu gdzieś między jednym a drugim punktem osi reprezentującej cechę.

Podobnie jak w deterministycznym modelu Guttmana zbiorowość osób odpowiadających na pytania wskaźnikowe jest dzielona na poziomy cechy ukrytej przez liczbę poprawnych odpowiedzi, których udzielili na zestaw py-

tań wskaźnikowych $X_1, \ldots X_k$. Ponieważ liczba ta może wynosić 0, w skalogramie Mokkena definiuje się *k+1* poziomów cechy ukrytej.

Operacja zliczania liczby poprawnych odpowiedzi ma swoje głębokie statystyczne uzasadnienie. Jak wykazał Mokken (1971: 124–129), liczba poprawnych odpowiedzi na pytania wskaźnikowe jest dla jego modelu – *de facto* – statystyką dostateczną poziomu cechy ukrytej **β**. Dzięki probabilizacji modelu reakcji na wskaźniki w sytuacji testowej udało się rozwiązać jeden z fundamentalnych problemów skalowania i statystycznie uzasadnić algorytm wyznaczania wartości cechy ukrytej na podstawie obserwowalnych jej wskaźników. Niestety, postulaty nakładane na funkcję reakcji są zbyt słabe, zbyt mało nakładają na tę funkcję ograniczeń, aby w równie zadowalający sposób rozwiązać pozostałe problemy skalowania, między innymi problem skalowalności.

## Stopień zgodności danych z modelem

Warunkiem skalowalności cechy ukrytej przez dowolny zestaw zmiennych wskaźnikowych jest możliwość ich uporządkowania ze względu na parametr **δ**. Oznacza to, iż przed rozpoczęciem skalowania powinno się sprawdzić, czy w zbiorze nie znajdują się wskaźniki zbędne, których poziom trudności jest taki sam. Standardowym narzędziem wspomagającym podejmowanie decyzji w takiej sprawie jest *test McNemara* dla hipotezy głoszącej, że dwie zmienne binarne mają w populacji tę samą średnią. Odrzucenie takiej hipotezy uzasadnia pozostawienie pary wskaźników w zestawie, nieodrzucenie – uzasadnia rezygnację z jednego z nich. W obu przypadkach uzasadnienie ma charakter statystyczny.

Wspomniany test McNemara odnosi się do założeń modelu. Ponadto do oceny stopnia skalowalności empirycznego zestawu wskaźników wykorzystuje się wspomniane wyżej konsekwencje podwójnej monotoniczności reakcji: pozytywną zależność wskaźników oraz monotoniczności macierzy $\mathbf{P}^{00}$ i $\mathbf{P}^{11}$. Kluczową rolę w parametryzacji pojęcia skalowalności odgrywają współczynniki H Loevingera.

## H Loevingera – współczynniki skalowalności

O ile test McNemara dotyczy wstępnego warunku podwójnej monotoniczności, istnienia porządku wskaźników względem **δ,** współczynniki Loevingera odnoszą się do stopnia, w jakim łączne rozkłady wskaźników odbiegają od za-

sady kumulatywności. Współczynnik zaproponowany przez Loevingera (Gilespe i in. 1987: 399) definiowany jest jako stosunek liczby „naruszeń" zasady kumulatywności zaobserwowanej w rozkładzie empirycznym i liczby „naruszeń" oczekiwanych w sytuacji, gdy wskaźniki są parami stochastycznie niezależne. Rozmiar „naruszenia" zasady kumulatywności reprezentują niezerowe częstości reakcji, które w modelu Guttmana zdarzać się nie mają prawa, czyli w miejscach, gdzie powinno się zaobserwować „strukturalne zera". Jest to zatem forma wyrażenia zgodności rozkładu empirycznego z oczekiwanym wedle modelu przy wykorzystaniu dwuzmiennowych rozkładów wskaźników.

Współczynnik $H_{ij}$ Loevingera dla wskaźników $X_i$, $X_j$, z których $X_j$ jest wskaźnikiem trudniejszym od $X_i$ , zatem $P(X_i=1) > P(X_j=1)$ określa się tak:

$$H_{ij} = 1 - \frac{P(X_i = 0, X_j = 1)}{P(X_i = 0)P(X_j = 1)} \tag{11}$$

Współczynnik $H_{ij}$ osiąga wartość 1gdy w rozkładzie występuje strukturalne zero, przyjmuje wartość 0, gdy wskaźniki są stochastycznie niezależne i jest ujemny, gdy wskaźniki są skorelowane negatywnie. W naszym przykładzie trzech wskaźników z tabeli 6 współczynniki Loevingera wynoszą odpowiednio: $H_{12} = 0{,}231$; $H_{13} = 0{,}091$; $H_{23} = -0{,}039$.

Współczynnik $H_{ij}$ daje się przedstawić w prostszej postaci:

$$H_{ij} = \frac{\operatorname{cov}(X_i, X_j)}{\operatorname{cov}(X_i, X_j)_{\max}} \tag{12}$$

gdzie licznik ułamka oznacza kowariancję zaobserwowaną w rozkładzie empirycznym, zaś mianownik jest graniczną wartością tej kowariancji osiąganą przy ustalonych rozkładach brzegowych obu wskaźników, a zatem przy założeniu, że w ich rozkładzie występuje strukturalne zero.

Suma współczynników $H_{ij}$ dla wszystkich par, w których występuje wskaźnik $X_i$ wyraża wedle Loevingerowi przydatność wskaźnika w skalowaniu kumulatywnym i jest zdefiniowana jako $H_i$:

$$H_i = \frac{\sum_{i=j+1, i\neq j}^{k} \operatorname{cov}(X_i, X_j)}{\sum_{i=j+1, i\neq j}^{k} \operatorname{cov}(X_i, X_j)_{\max}} \tag{13}$$

Sumując przydatności $H_i$ po wszystkich wskaźnikach otrzymujemy pośrednią miarę zgodności łącznego rozkładu zestawu wskaźników z założeniem podwójnej monotoniczności reakcji:

$$H = \frac{\sum_{i=j+1}^{k} \sum_{j=1}^{k-1} \operatorname{cov}(X_i, X_j)}{\sum_{i=j+1}^{k} \sum_{j=1}^{k-1} \operatorname{cov}(X_i, X_j)_{\max}} \tag{14}$$

Wedle Mokkena (Mokken, Lewis 1982: 417–430) współczynniki Loevingersa pozwalają uzasadnić decyzję o uznaniu za skalowalny cały zestaw wskaźników, gdy $H > 0{,}3$ lub o zrezygnowaniu z pojedynczych wskaźników dla których $H_i < 0{,}3$. Wybór tych akurat wartości wynikał, zdaniem Mokkena, z „doświadczenia empirycznego", co trudno uznać za uzasadnienie zadowalające.

## Skalogram Mokkena – podsumowanie

Sprawdzimy teraz w jaki sposób model Mokkena radzi sobie z najważniejszymi wyzwaniami stawianymi przed dobrym modelem skalowania.

**I. Problem skalowalności**

Zalążkowe kryteria, często typu *ad hoc* pozwalają uznać zestaw wskaźników określonych w pewnej zbiorowości za nieskalowalny, jeśli zdarzy się jedna z dwóch sytuacji, współczynnik Lovingera dla całego zestawu będzie niższy niż 0,3 (według Mokkena) albo, gdy macierze $\mathbf{P}^{11}$ i $\mathbf{P}^{00}$ będą zawierały wartości zbyt odległe od oczekiwanych, przy czym nie wiadomo, co to znaczy „zbyt odległe". Trudno uznać powyższe kryteria rozstrzygania za dobrze uzasadnione, a ponadto, jak wskazuje przykład tabeli 6, oba te warunki są względnie od siebie niezależne.

**II. Problem liczby wymiarów cechy ukrytej i relacji między nimi**

Podobnie jak w modelu Guttmana, w skalogramie Mokkena nie ma procedury pozwalającej rozstrzygać tę kwestię.

**III. Czy wszystkie wskaźniki są potrzebne?**

Ocena własności pojedynczych wskaźników jest przeprowadzana na dwa sposoby, z których pierwszy ma w pełni statystyczny charakter.

Standardowy test dla identyczności populacyjnych proporcji dla zmiennych binarnych (test McNemara) pozwala w zestawie wskaźników wyeliminować te, które są w nim zbędne, a decyzję przekonywująco, bo statystycznie, uzasadnić.

Słabsze podstawy statystyczne ma decyzja o eliminacji z zestawu wskaźnika, dla którego współczynnik Loevingera $H_i$ przyjmuje wartość niższą niż 0,3. Niektórzy autorzy jako uzasadnienie decyzji o niekumulatywności rozkładów łącznych, w których występuje wskaźnik $X_i$ proponują używać statystyki testującej hipotezę, że współczynnik Loevingera $H_i$ dla tego wskaźnika ma w populacji wartość 0, przeciwko hipotezie, że tak nie jest[11]. Zauważmy jednak, że w ten sposób testowana jest hipoteza o niezależności stochastycznej wskaźników, a nie o zgodności ich rozkładu z konsekwencjami założenia podwójnej monotoniczności.

**IV. Jakie są własności diagnostyczne poszczególnych wskaźników?**

Test McNemara pozwala wykryć wskaźniki o identycznych włąsnościach diagnostycznych, o tym samym poziomie trudności. Innych parametrów własności diagnostycznych pytań wskaźnikowych skalogram Mokkena nie przewiduje

**V. Jak skalować – funkcja agregująca profile**

Skalogram Mokkena jest w skalowaniu kumulatywnym etapem przełomowym, gdyż jako pierwszy miał statystyczne uzasadnienie dla postaci swojej funkcji agregującej informacje zawarte w empirycznym rozkładzie profili reakcji w procedurze wyznaczania wartości zmiennej ukrytej. Mokken wykazał, że liczba poprawnych odpowiedzi jest statystyką dostateczną dla skalowanej cechy ukrytej. Jest to minimalny wymóg stawiany wszelkim procedurom estymacyjnym i dzięki jego spełnianiu można powiedzieć, że w modelu Mokkena oszacowuje się częstości rozkładu cechy ukrytej $\boldsymbol{\beta}$.

Skalogram Mokkena jest uogólnieniem modelu Guttmana, gdyż spełnia zasadę współmierności, kumulatywności reakcji i lokalnej niezależności reakcji. Jednakże w modelu Mokkena funkcja reakcji obiektu na wskaźnik określana przez założenie podwójnej monotoniczności ma rzeczywiście probabilistyczny charakter. Dzięki temu kilka problemów, z którymi model Guttmana sobie nie radził, uzyskało w skalogramie Mokkena zadowalające rozwiązanie.

[11] Argumentem przemawiającym za tym rozwiązaniem jest zbieżność statystyki z prób do rozkładu normalnego (patrz Komorek 2006).

Nieparametryczny charakter modelu, dzięki słabszym założeniom od modeli parametrycznych (patrz następny rozdział), pozwala na zaakceptowanie jako zgodnych z modelem obszernego zbioru empirycznych funkcji reakcji o niekoniecznie regularnym charakterze. Podobnego charakteru funkcje reakcji, tym razem sparametryzowane, będą rozważane w modelach Rascha w następnym rozdziale.

# Model Rascha

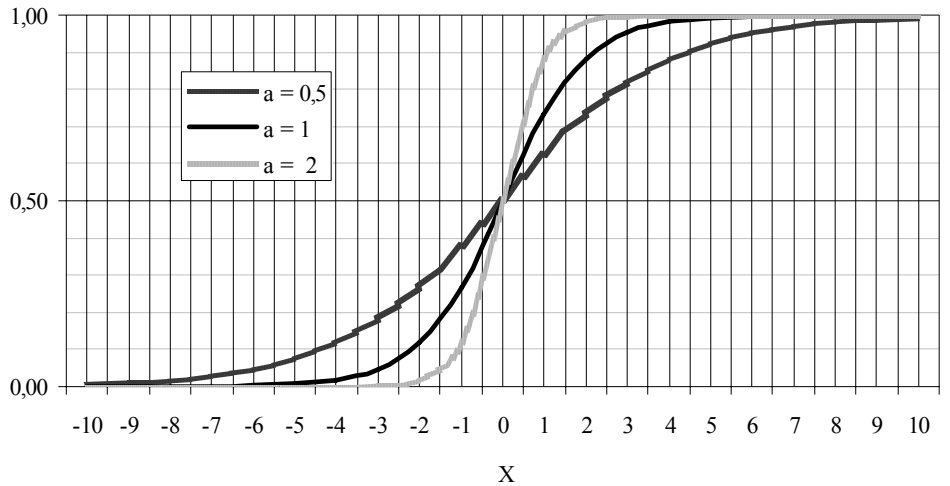
Model skalowania Rascha (1980) jest z jednej strony szczególnym przypadkiem skalogramu Mokkena, gdyż spełnia wszystkie jego założenia, z drugiej strony jest konstruktem od podstaw probabilistycznym i otwartym, dzięki czemu oferuje to, czego skalogramy Mokkena i Guttmana nie miały – możliwość szacowania odległości między poziomami cechy ukrytej. Możliwość ta, podobnie jak spełnianie założeń współmierności, punktu przecięcia poziomu 0,5 i podwójnej kumulatywności jest efektem założenia, że funkcja reakcji jest funkcją logistyczną definiowaną tak:

$$f(\mathrm{x}) = \frac{e^{a\,\mathrm{x}}}{1 + e^{a\,\mathrm{x}}} \tag{15}$$

Funkcja ta jest symetryczna względem zera i przyjmuje wartości z przedziału (0, 1), do którego granic zmierza asymptotycznie. Kąt nachylenia krzywej logistycznej zależy od wartości współczynnika $a \neq 0$. Funkcja osiąga wartość 0,5 dla argumentu równego 0, przyjmuje wartości poniżej 0,5 dla argumentów ujemnych, a powyżej 0,5 dla dodatnich.

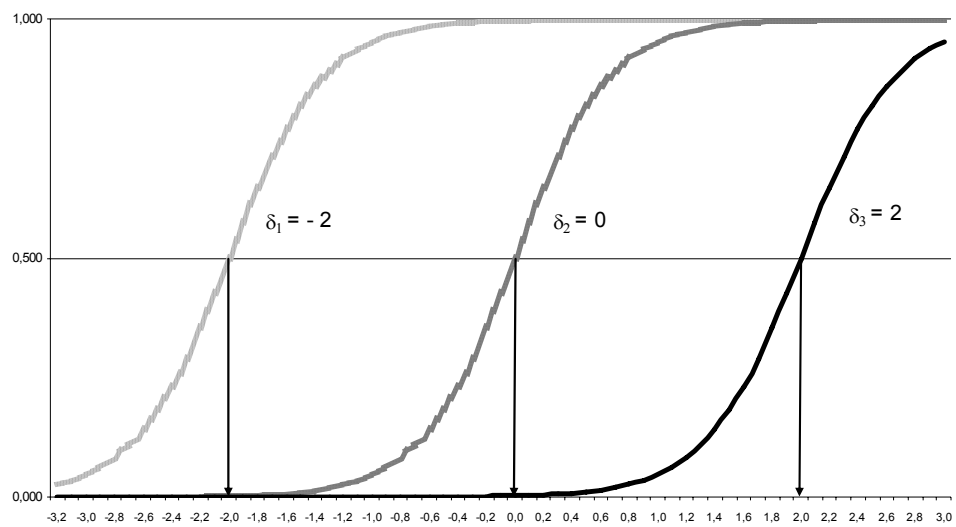
Logistyczna funkcja argumentów rzeczywistych może zostać zinterpretowana jako funkcja reakcji, gdyż jej wartości nie przekraczają granic, w których prawdopodobieństwo musi się znajdować. Wystarczyło teraz przyjąć, że pozioma oś wykresu funkcji logistycznej reprezentuje różnicę między poziomem umiejętności osoby $\varpi_{\mathrm{v}}$ i poziomem trudności pytania $X_i$, czyli między $\beta_{\mathrm{v}}$ i $\delta_i$, aby zdefiniować funkcję reakcji na wskaźnik $X_i$, określoną przez:

$$P_{\beta\delta}\left(X_i = 1 \,|\, \boldsymbol{\beta} = \beta_v\right) = \frac{e^{(\beta_v - \delta_i)}}{1 + e^{(\beta_v - \delta_i)}} \tag{16}$$




**Rysunek 6.** Wykres trzech funkcji logistycznych różniących się parametrem *a*

Funkcja ta definiuje najprostszy model Rascha, zwany 1PL, w którym stała *a* jest równa 1, co oznacza, że krzywe reakcji wszystkich zmiennych wskaźnikowych mają takie samo nachylenie. Oto przykład trzech takich funkcji dla wskaźników o poziomach trudności –2, 0 oraz +2:




**Rysunek 7.** Krzywe reakcji na pytania testowe $X_1$, $X_2$ i $X_3$ o trudności –2, 0 i +2.

Pamiętając, że oś pozioma charakteryzuje także poziom umiejętności testowanych osób, zobaczymy, że prawdopodobieństwo poprawnej odpowiedzi na

pytanie $X_1$ powyżej 0,5 mają osoby, których poziom umiejętności wynosi co najmniej –1, zaś w wypadku pytania wskaźnikowego $X_3$ poziom ten musi wynosić co najmniej +2. Z wykresów powyżej widać również, że wskaźniki z logistyczną funkcją reakcji spełniają postulat podwójnej monotoniczności z modelu Mokkena. Widać również, że wybór jednostek wyrażających trudność wskaźników i poziom umiejętności osób jest arbitralny – kształt funkcji reakcji nie zmienia się przy równoległych liniowych przekształceniach obu wielkości, co oznacza także, że punkt zerowy osi poziomej jest wybierany arbitralnie.

## Warianty modelu skalowania Rascha

Definicja modelu skalowania Rascha zawiera dwa elementy wspólne obu modelom omawianym wcześniej:

- założenie o współmierności cechy ukrytej i parametrów wskaźników,
- oraz założenie lokalnej niezależności reakcji.

Tym, co specyficzne dla modeli Rascha, jest sposób definiowania funkcji reakcji.

W najprostszym modelu typu 1PL funkcja ta zdefiniowana przez:

$$P_{\beta\delta}\left(X_i = 1 \mid \boldsymbol{\beta} = \beta_v\right) = \frac{e^{(\beta_v - \delta_i)}}{1 + e^{(\beta_v - \delta_i)}} \tag{17}$$

zakłada *implicite* współmierność parametrów osób i pytań i gwarantuje spełnienie postulatu podwójnej monotoniczności Mokkena.

Rozszerzeniem modelu 1PL jest model 2PL, w którym kąt nachylenia funkcji reakcji nie musi być jak poprzednio identyczny dla wszystkich wskaźników, lecz różny, w zależności od tego, jaką wartość dla wskaźnika przybiera odpowiedzialny za to parametr *a*:

$$P_{\beta\delta}\left(X_i = 1 \mid \boldsymbol{\beta} = \beta_v\right) = \frac{e^{a_i(\beta_v - \delta_i)}}{1 + e^{a_i(\beta_v - \delta_i)}} \tag{18}$$

Gdy parametr $a_i$ jest równy 1, model 2PL staje się modelem 1PL. Nazwa modelu Rascha zawiera informację o tym, ile parametrów wskaźnika występuje w funkcji reakcji – w modelu 1PL występuje tylko parametr trudności pytania ($\delta_i$), w modelu 2PL – dwa parametry $a_i$ oraz $\delta_i$ . W praktyce skalowania osiągnięć szkolnych stosuje się trójparametryczny model 3PL, w którym funkcja reakcji jest tylko częściowo logistyczna:

$$P_{\beta\delta}\left(X_i = 1 \mid \boldsymbol{\beta} = \beta_v\right) = c_i + \left(1 - c_i\right)\frac{e^{a_i(\beta_v - \delta_i)}}{1 + e^{a_i(\beta_v - \delta_i)}} \tag{19}$$

Parametr $c_i$ jest tak zwanym „współczynnikiem zgadywania" i reprezentuje jednakowe dla wszystkich osób, lecz różne dla różnych pytań prawdopodobieństwo odgadnięcia poprawnej odpowiedzi. Model 3PL zakłada zatem, że pewna frakcja poprawnych odpowiedzi, dokładnie $c_i$, nie ma związku z umiejętnościami osób ani z trudnością pytania, lecz została z prawdopodobieństwem $c_i$ „wylosowana" przez odpowiadających. Model 3PL nie jest w pełni „logistyczny", stąd przy estymacji jego parametrów pojawiają się kłopoty, których nie ma w wypadku estymacji parametrów modeli 1PL i 2PL.

## Własności modelu Rascha

Z powodu probabilistycznej konstrukcji modelu Rascha jego elementy opisywane są w języku statystyki opisowej i inferencyjnej, zaś rozwiązywanie podstawowych i szczegółowych problemów skalowania staje się rozwiązywaniem niekiedy standardowych, a niekiedy nietypowych i trudnych problemów estymacji bądź weryfikacji hipotez statystycznych. Pozwala to na sięganie do zasobów teoretycznych współczesnej statystyki i skorzystanie z techniki komputerowej (także metod symulacyjnych), a dzięki temu wypracowanie rozwiązań dobrze statystycznie uzasadnionych. Omówimy je dla najprostszego wariantu skalowania Rascha – modelu 1PL. We wszystkich modelach Rascha estymacja ich parametrów, a także weryfikacja hipotez na ich temat odbywa się przy wykorzystaniu funkcji wiarygodności parametru (zob. Agresti 2002, Eliason 1993, Pawłowski 1980, Silvey 1978).

### Estymacja parametrów

W modelu Rascha 1PL należy oszacować wartości parametrów pytań i osób na podstawie łącznego rozkładu binarnych wskaźników ($X_1$, $X_2$, $X_3$, ..., $X_i$, ..., $X_k$) traktowanego jako realizacja $k$-wymiarowej zmiennej losowej, której wartościami są profile reakcji. Rozkład prawdopodobieństwa tej zmiennej określany jest przez dwa wektory o wartościach rzeczywistych: wektor $\boldsymbol{\delta}$ „poziomów trudności wskaźników" oraz wektor $\boldsymbol{\beta}$ „poziomów umiejętności osób".

Szacowanie parametrów jakiegokolwiek modelu probabilistycznego musi zostać poprzedzone znalezieniem dla niego statystyki, najlepiej dostatecznej. Dla modelu 1PL nie jest to trudne, okazuje się bowiem, że:

(i) Jeśli założyć, że dany jest wektor **β** „poziomów umiejętności osób", statystyką dostateczną dla poziomu trudności $\delta_i$ pytania wskaźnikowego $X_i$ jest liczba poprawnych odpowiedzi na to pytanie:

$$s_i = \sum_{v=1}(x_{iv}) \tag{20}$$

(ii) Jeśli założyć, że dany jest wektor **δ** „poziomów trudności wskaźników", statystyką dostateczną dla poziomu umiejętności $\beta_v$ osoby $\omega_v$ jest liczba pytań, na które odpowiedziała poprawnie:

$$r_v = \sum_{i=1}(x_{iv}) \tag{21}$$

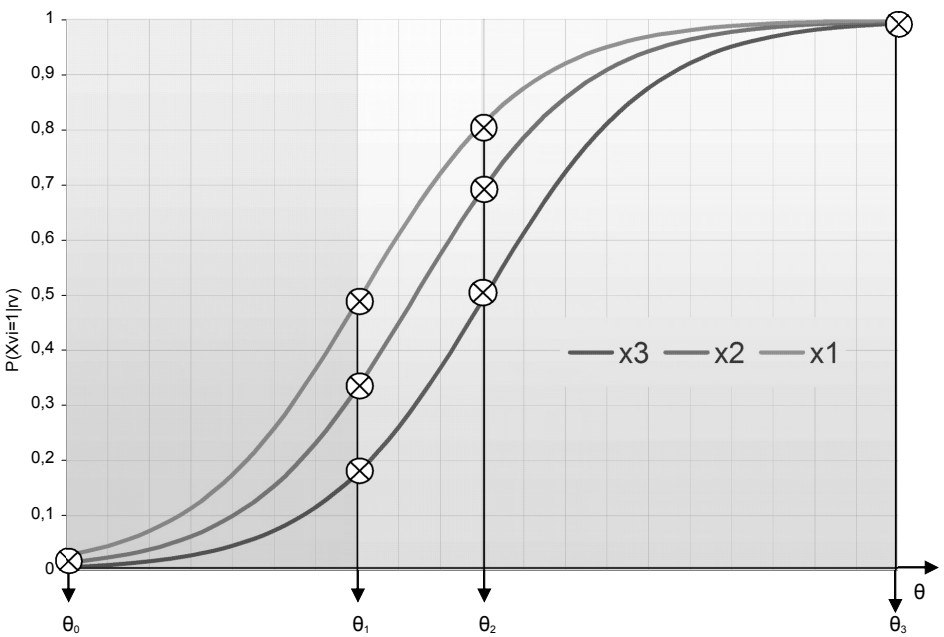
Sens obu twierdzeń jest prosty: osobom, które udzieliły tyle samo poprawnych odpowiedzi, zostanie przydzielony ten sam poziom umiejętności, zaś odsetek poprawnych odpowiedzi na pytanie jednoznacznie określi poziom jego trudności. Stanie się tak ze znanego statystykom powodu – estymator największej wiarygodności jest zawsze funkcją statystyki dostatecznej (zob. Silvey 1978).

Wynik jest nieco zaskakujący – funkcja agregująca profile reakcji w poziomy cechy ukrytej jest taka sama jak w modelach Guttmana i Mokkena. Trzeba zwrócić uwagę na warunkowy charakter obu twierdzeń: mówią one w gruncie rzeczy o warunkowej procedurze estymacji, która zakłada, że jeden rodzaj parametrów jest już oszacowany i poszukujemy oszacowania dla pozostałych. Taka warunkowa procedura estymacyjna nie jest jedyną stosowaną w modelach Rascha. Zanim przedstawimy następne, spróbujmy przedstawić intuicje, które kryją się za twierdzeniami o dostateczności[12].

Osoby, które zebrały tę samą liczbę punktów, mogą mieć różny poziom umiejętności, lecz w grupie osób, które odpowiedziały na tyle samo pytań, ich proporcje zależą od tego na które z pytań odpowiedziały poprawnie – na łatwe czy też na trudne. Każda osoba zajmuje jakieś miejsce na osi rzeczywistej reprezentującej jej poziom umiejętności, a to oznacza, że prawdopodobieństwo dowolnego profilu reakcji tej osoby (w tym profilu dającego stałą

[12] Dowód obu powyższych twierdzeń jest prosty (zob. Komorek 2006) i niektóre specjalistyczne monografie statystyki IRT pomijają go z tego powodu.

sumę punktów) zależy wyłącznie od tego, z których pytań pochodzą punkty za odpowiedzi poprawne – z łatwych czy z trudnych. Model Rascha „pozwala" każdej osobie o dowolnym poziomie umiejętności odpowiedzieć poprawnie na wszystkie trzy pytania testu, lecz prawdopodobieństwo takiego zdarzenia jest bardzo niskie dla osób o niskim poziomie umiejętności i bardzo wysokie dla osób o poziomie umiejętności bardzo wysokim, jak widać na rysunku 8 (Komorek 2006) poniżej.




**Rysunek 8.** Funkcje reakcji na pytania $X_1$, $X_2$, $X_3$ w modelu Rascha 1PL – przykład

Rysunek ten ilustruje również powód, dla którego statystyką dostateczną dla trudności pytania jest liczba osób, które odpowiedziały poprawnie. Logistyczna funkcja reakcji spełnia warunek podwójnej monotoniczności, zatem reakcje na wskaźniki są między sobą pozytywnie zależne, więc ich porządek jest wyznaczany przez częstości odpowiedzi poprawnych. Ponieważ na każdy wskaźnik odpowiedzi udzielają te same osoby, bez względu na rozkład ich poziomów umiejętności, punkt przecięcia funkcji reakcji wskaźnika z poziomem 0,5 zależy wyłącznie od wartości parametru trudności pytania.

Dzięki twierdzeniom o dostateczności statystyk dla obu rodzajów parametrów możliwa jest ich estymacja wykorzystująca funkcję wiarygodności.

W skalowaniu kumulatywnym, czy też ogólniej, w teorii reakcji (IRT), w estymacji parametrów maksymalizuje się różne warianty funkcji wiarygodności. Różnią się od siebie rodzajem założeń na temat parametrów i danych, których używają do ich estymacji oraz postacią funkcji, którą się maksymalizuje (zob. Andrich 1985, 1988; Baker, Kim 2004; Fischer, Molenaar 1996; Linden, Hambleton 1977). Każdy z nich ma jakieś wady i zalety[13].

Wektory parametrów modelu Rascha można estymować osobno, a wówczas mamy do czynienia z estymacją typu CML (ang. *Conditional Maximum Likelihood – CML*). Twierdzenia o dostateczności statystyk $s_i$ oraz $r_v$ dotyczą właśnie takiej procedury. Jej wadą „konstrukcyjną" jest właśnie konieczność przyjmowania założeń o jednym z wektorów w celu oszacowania wartości drugiego, co czyni wnioskowanie cyklicznym.

Gdy procedura usiłuje wyznaczyć jednocześnie estymatory obu wektorów parametrów modelu mamy do czynienia z estymacją typu JML (ang. *Joint Maximum Likelihood – JML*). Nie jest polecana przez autorytety IRT[14], gdyż źle radzi sobie ze „źle uwarunkowanymi" (nieregularnymi) rozkładami cechy ukrytej i *de facto* nie pozwala oszacować skrajnych poziomów jej wartości, odpowiadających profilom z minimalną lub maksymalną liczbą poprawnych odpowiedzi.

W procedurze MML (ang. *Marginal Maximum Likelihood – MML*) zamiast estymować rozkład cechy ukrytej przyjmuje się, że jest on jakiś – najczęściej, że jest normalny – i przy tym założeniu poszukuje się najbardziej prawdopodobnych wartości parametrów. Podobne założenia przyjmuje procedura BME (ang. *Bayesian Modal Estimator – BME)*.

W projekcie PISA parametry osób szacuje się metodą WLE (ang. *Weighted Likelihood Estimator – WLE)* od nazwiska autora nazywana metodą Warma (Warm 1989). Pozwala ona efektywnie szacować także skrajne wartości cechy ukrytej.

## Przykład skalowania 1PL

Tabela 9 zawiera wynik skalowania 1PL zestawu trzech dychotomicznych wskaźników używanego w poprzednich przykładach.

---

[13] Przegląd własności procedur estymacyjnych można znaleźć w Kim, Nicewander (1993).

[14] Jak na przykład Verhelst (2007).

**Tabela 9.** Skalowanie cechy ukrytej za pomocą trzech dychotomicznych wskaźników za pomocą modelu Rascha 1PL

| | | Model Rasch 1PL | | | | | | Oczekiwane częstości profili | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Funkcje reakcji | | | $\Sigma_j P(X_1=x_1, X_2=x_2, X_3=x_3 \mid \boldsymbol{\beta}=\beta_v)P(\boldsymbol{\beta}=\beta_v)$ | | | | | | | |
| Rozkład **β** | | Profile Guttmana | | | $P(X_i=1 \mid \boldsymbol{\beta}=\beta_v)$ | | | 0,37 | 0,12 | 0,10 | 0,13 | 0,08 | 0,07 | 0,08 | 0,06 |
| $\beta_j$ | $P(\boldsymbol{\beta}=\beta_v)$ | $X_1$ | $X_2$ | $X_3$ | $X_1$ | $X_2$ | $X_3$ | **111** | **110** | **100** | **000** | **101** | **011** | **010** | **001** |
| 3,00 | 0,30 | 1 | 1 | 1 | 0,98 | 0,97 | 0,96 | 0,91 | 0,04 | 0,00 | 0,00 | 0,03 | 0,02 | 0,00 | 0,00 |
| –0,05 | 0,35 | 1 | 1 | 0 | 0,66 | 0,62 | 0,53 | 0,22 | 0,19 | 0,12 | 0,06 | 0,13 | 0,11 | 0,10 | 0,07 |
| –0,89 | 0,30 | 1 | 0 | 0 | 0,45 | 0,41 | 0,33 | 0,06 | 0,12 | 0,18 | 0,22 | 0,09 | 0,07 | 0,15 | 0,11 |
| –3,00 | 0,05 | 0 | 0 | 0 | 0,09 | 0,08 | 0,06 | 0,00 | 0,01 | 0,08 | 0,79 | 0,00 | 0,00 | 0,07 | 0,05 |
| | | Trudność pytań | | | ⇐ MLE ⇑⇑ | | | Empiryczny rozkład profili reakcji | | | | | | | |
| | | $\delta(X_1)$ | $\delta(X_2)$ | $\delta(X_3)$ | | | | $p_{111}$ | $p_{110}$ | $p_{100}$ | $p_{000}$ | $p_{101}$ | $p_{011}$ | $p_{010}$ | $p_{001}$ |
| | | –0,71 | –0,53 | –0,18 | | | | 0,30 | 0,20 | 0,10 | 0,05 | 0,10 | 0,05 | 0,10 | 0,10 |
| | | | | | | | | Stopień zgodności empirycznego rozkładu profili z częstościami oczekiwanymi | | | | | | | |
| | | Różnica między empirycznymi częstościami profili w próbie a częstościami oczekiwanymi przez model | | | | | | *–0,07* | *0,08* | *0,00* | *–0,08* | *0,02* | *–0,02* | *0,02* | *0,04* |

Estymację wykonano metodą JML przy użyciu dedykowanego języka oprogramowania ICL (Kryniewski 2007). Liczba i rozkład poziomów cechy ukrytej jest taka sama jak w poprzednich modelach skalowania (patrz tabele 5, 6), lecz tym razem możemy interpretować odległości między nimi. Nie są one równe, jak poprzednio. Zwróćmy również uwagę, że parametry trudności pytań są ujemne, co wynika z faktu (patrz tabele 5–6), że brzegowe proporcje poprawnych odpowiedzi na każde z pytań wskaźnikowych przekraczały 0,5.

W porównaniu z poprzednimi modelami warunkowe prawdopodobieństwa reakcji są bardziej zróżnicowane niż poprzednio, lecz zgodnie z oczekiwaniami spełniają warunek podwójnej monotoniczności Mokkena. W rezultacie w warstwach definiowanych przez poziom cechy ukrytej, w każdym z czterech wierszy prawej strony tabeli warunkowe łączne prawdopodobieństwa profili reakcji są najwyższe dla profili z klasy dopuszczalnych w modelu Guttmana. Dla przykładu, osoba należąca do klasy ukrytej, dla której $\beta(\varpi) = -3{,}0$, czyli do klasy o najniższym poziomie umiejętności, może udzielić odpowiedzieć poprawnej odpowiedzi na najłatwiejsze pytanie (profil 100) albo nawet na dwa najłatwiejsze pytania (profil 110), lecz prawdopodobieństwa takich zdarzeń, wedle modelu są bardzo niskie i wynoszą 0,08 i 0,01 odpowiednio. Najbardziej prawdopodobnym profilem reakcji dla osób z tej klasy jest profil 000, którego prawdopodobieństwo (warunkowe!) model szacuje na 0,79.

Wartości obu wektorów parametrów są podawane z dokładnością do liniowego przekształcenia – wybór miary odległości i punktu zerowego jest arbitralny. Oznacza to, że w praktycznych zastosowaniach modelu Rascha zachodzi konieczność kalibracji wartości cechy ukrytej, a więc i poziomów trudności pytań. Procedury estymacyjne przyjmują niekiedy (dla wygody), że suma parametrów wynosi zero, co skłania do interpretowania poziomów wartości cechy ukrytej jako odchyleń od ich średniej w badanej zbiorowości.

## Przykład skalowania 2PL

Użyjemy tego samego, co poprzednio zestawu trzech wskaźników i wyskalujemy poziomy cechy ukrytej oraz trudności pytań wskaźnikowych przy użyciu modelu 2PL, w którym funkcja reakcji każdego ze wskaźników może mieć inne nachylenie. Oznacza to, że oprócz poziomu trudności wskaźnika trzeba będzie oszacować parametr odpowiedzialny za tempo przyrastania je-

**Tabela 10.** Skalowanie cechy ukrytej za pomocą trzech dychotomicznych wskaźników za pomocą modelu Rascha 2PL

| Model Rasch 2PL | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Cecha ukryta | Profil reakcji | Funkcje reakcji $P(X_i=1\|\boldsymbol{\beta}=\beta_v)$ | | | | Częstości profili reakcji | | |
| j | $\beta_v$ | $\langle x_1, x_2, x_3\rangle$ | $X_1$ | $X_2$ | $X_3$ | Liczba poprawnych odpowiedzi | Oczekiwane | W próbie | Różnica |
| 1 | –3,00 | **000** | 0,15 | 0,20 | 0,30 | 0 | 0,07 | **0,05** | 0,02 |
| 2 | –0,95 | **100** | 0,48 | 0,48 | 0,49 | 1 | 0,04 | **0,10** | –0,06 |
| 3 | –1,28 | **010** | 0,41 | 0,43 | 0,46 | 1 | 0,04 | **0,10** | –0,06 |
| 4 | –1,90 | **001** | 0,30 | 0,34 | 0,40 | 1 | 0,05 | **0,10** | –0,05 |
| 5 | 0,50 | **110** | 0,74 | 0,70 | 0,63 | 2 | 0,11 | **0,20** | –0,09 |
| 6 | –0,16 | **101** | 0,63 | 0,61 | 0,57 | 2 | 0,04 | **0,10** | –0,06 |
| 7 | –0,49 | **011** | 0,57 | 0,55 | 0,53 | 2 | 0,02 | **0,05** | –0,03 |
| 8 | 3,00 | **111** | 0,96 | 0,92 | 0,82 | 3 | 0,63 | **0,30** | 0,33 |
| Współczynnik trudności $\delta(X_i)$ | | | –0,83 | –0,70 | –0,33 | | | | |
| Współczynnik dyskryminacji $a_i$ | | | 0,80 | 0,64 | 0,39 | | | | |

go funkcji reakcji, nazywany dalej współczynnikiem dyskryminacji. Wyniki przedstawia tabela 10.

Obliczenia zostały wykonane metodą JML przy użyciu tego samego co poprzednio języka ICL.

Gdyby spojrzeć tylko na wiersze oznaczone szarym tłem, otrzymane rozwiązanie byłoby podobne do poprzedniego i przypominało model skalogram Mokkena. Warunkowe prawdopodobieństwa reakcji dla poziomów cechy ukrytej oznaczonych szarym tłem, odpowiadających profilom dopuszczalnym w modelu Guttmana, spełniają postulat podwójnej monotoniczności.

Jednakże model różni się od poprzedniego znacznie. Dzięki temu, że każda funkcja reakcji ma swoje nachylenie, którego współczynnik nazywany jest współczynnikiem dyskryminacji pytania, można rozwiązać środkami statystycznymi problem wyznaczenia wartości cechy ukrytej dla profili niedopuszczalnych w modelu Guttmana, czyli dla osób reagujących niezgodnie z założeniem kumulatywności. W modelu 2PL wszystkie profile reakcji mają „swoją" wartość cechy ukrytej.

Na komentarz zasługują wartości współczynników dyskryminacji. Dla wszystkich pytań są one niższe od 1, co oznacza, że tempo zmian prawdopodobieństwa reakcji na pytania wskaźnikowe jest niższe od tego, które miałaby funkcja reakcji utworzona z dystrybuanty rozkładu normalnego. W tym sensie przykładowe wskaźniki mają własności dyskryminacyjne niższe niż „normalne".

Znaczenie i doniosłość współczynników dyskryminacji w praktyce skalowania omówimy dalej, przedtem skomentujmy rozbieżność między częstościami profili reakcji oczekiwanymi wedle wyestymowanego modelu 2PL i empirycznymi. Zgodnie z zapowiedzią estymatory typu JML nie radzą sobie ze skrajnymi poziomami cechy ukrytej – częstości odpowiadających im profili (**000** i **111** ) są przeszacowane, a częstość drugiego z nich wręcz dwukrotnie.

## Funkcja informacyjna wskaźnika

W skalowaniu Rascha pojedynczy wskaźnik może być charakteryzowany za pomocą jednego lub kilku parametrów, w zależności od tego, jaką postać ma model. Bez względu na to, ile tych parametrów jest, w praktyce konstruowania testów najważniejsze są dwie własności:

1) wartość dyskryminacyjna wskaźnika, czyli jego skuteczność w diagnozowaniu poziomu cechy ukrytej,

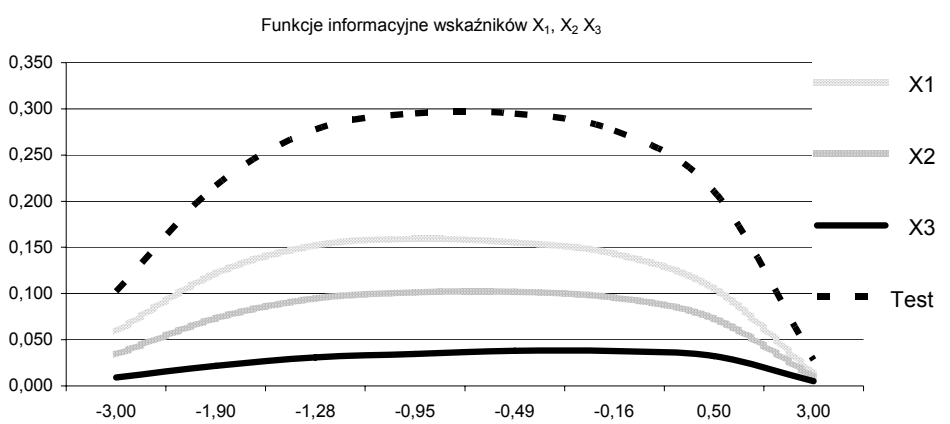
2) miejsce na osi reprezentującej poziom cechy ukrytej, nad którą wskaźnik najlepsze własności diagnostyczne.

Wartość dyskryminacyjna wskaźnika dychotomicznego reprezentowana jest przez współczynnik *a* z modelu 2PL. Jego wielkość wyznacza tak zwaną wartość informacyjną wskaźnika, czyli sumę kwadratów błędów estymacji poziomów cechy ukrytej. Zależność precyzji szacowania cechy ukrytej przez wskaźnik zależny od jej poziomu przedstawia tak zwana funkcja informacyjna wskaźnika, która dla wskaźników dychotomicznych ma postać:

$$I\left(X_i,\boldsymbol{\beta}\right)=I_i\left(\boldsymbol{\beta}\right)=a_i^2P\left(X_i=1\,|\,\boldsymbol{\beta}\right)\left(\left(1-P\left(X_i=1\,|\,\boldsymbol{\beta}\right)\right)\right) \qquad (22)$$

Funkcja ta osiąga maximum dla poziomu cechy ukrytej odpowiadającej poziomowi trudności wskaźnika $X_i$, czyli dla punktu $\boldsymbol{\beta} = \delta(X_i)$, w którym prawdopodobieństwo poprawnej dopowiedzi na pytanie wskaźnikowe wynosi 0,5. Z zasady współmierności partametrów osób i parametrów pytań wynika, że dla każdego pytania wskaźnikowego taki punkt istnieje. Wobec tego wartość maksimum funkcji informacyjnej dychotomicznego wskaźnika zależy wyłącznie od wartości współczynnika dyskryminacji pytania $a_i$.

W naszym przykładzie z poprzedniej tabeli funkcje informacyjne wskaźników mają następujący przebieg:




**Rysunek 9.** Funkcje informacyjne wskaźników z tabeli 10

Jak widać, wskaźniki z naszego przykładu, przyznajmy, złośliwie zaprojektowanego, maja płaskie krzywe informacyjne, co zawdzięczają niskim współczynnikom dyskryminacji. Ponadto, funkcje osiągają maksymalne war-

tości nad punktami odpowiadającymi ujemnym wartościom cechy ukrytej, nadają się zatem raczej do diagnozowania umiejętności poniżej poziomu przeciętnego (co nie jest zaskakujące, jeśli pamiętać, że na najłatwiejsze pytanie ($X_1$) poprawnej odpowiedzi udzieliło aż 70% zbiorowości, na najtrudniejsze ($X_I$) aż 55%.

Wartość diagnostyczna pytań jest zatem niewielka. W konsekwencji niewielka jest również maksymalna wartość funkcji informacyjnej całego testu, która dzięki założeniu o lokalnej niezależności reakcji jest sumą wartości funkcji informacyjnych wszystkich wskaźników:

$$I(\beta) = \sum_{i=1}^{k} I_i(\beta) \tag{23}$$

Najniższy poziom błędu standardowego szacowania cechy ukrytej, największą precyzję oszacowań, nasz zestaw testowy osiąga, podobnie jak jego wskaźniki, dla niskich poziomów umiejętności, reprezentowanych przez ujemne wartości **β**. Dzieje się tak, gdyż błąd standardowy estymacji poziomu wartości cechy ukrytej jest równy:

$$SE(\beta) = \frac{1}{\sqrt{I(\beta)}} \tag{24}$$

Wynika stąd i z addytywności funkcji informacyjnej zestawu testowego, że aby test miał wysoką wartość diagnostyczną w całym zakresie cechy ukrytej, powinien składać się z wielu wskaźników, a wskaźniki te swoimi poziomami trudności (punktami, w których są najbardziej precyzyjne) powinny pokrywać możliwie szeroki zakres poziomów skalowanej cechy. Odpowiada to zdroworozsądkowemu przekonaniu, że dobry test powinien mieć wiele pytań o zróżnicowanym poziomie trudności. Aby tak się stało, potrzebny jest jeszcze jeden warunek, którego spełnienie nie zależy zazwyczaj od autorów testu – w badanej zbiorowości muszą się znaleźć osoby, których poziomy umiejętności korespondują z poziomem trudności pytań.

## Skalowalność

Dzięki probabilistycznym podstawom modelu Rascha jego parametry wyznaczane są przy użyciu standardowych lub specyficznych technik estymacyjnych, a używane przy tym estymatory są oceniane z punktu widzenia standardowych w statystyce kryteriów jakości: obciążenia, efektywności,

wielkości błędu oszacowań. Użyteczność zwłaszcza tego ostatniego kryterium ilustrowały ostatnie przykłady.

Probabilistyczny charakter modelu powoduje również, iż skalowanie Rascha może sięgać do bogatych zasobów technik weryfikacji hipotez posługujących się statystyką ilorazu wiarygodności, co ułatwia zwłaszcza wykładniczy charakter funkcji reakcji. Parametryczny charakter modeli Rascha czyni testowanie hipotez na temat własności jego elementów, osób lub wskaźników, zadaniem rutynowym. Do standardowych w statystyce inferencyjnej posługującej się funkcjami wiarygodności należy również ocena zgodności rozkładu empirycznego z przewidywaniami modelu. Maksymalizacja funkcji wiarygodności pozwala bowiem zasadnie wybrać odpowiedź na pytanie o wartości parametrów (oszacowane przecież najlepiej jak można), nie gwarantuje jednak, że ta najlepsza z możliwych parametryzacja pozwala dobrze odtwarzać empiryczny rozkład zmiennych obserwowalnych. Spektakularną porażką w tym względzie okazał się model 2PL z tabeli 10.

## Modele Rascha – podsumowanie

Rodzina modeli Rascha spełnia wszystkie postulaty stawiane dobremu modelowi skalowania. Dzięki potraktowaniu wyników testu jako realizacji wielowymiarowej zmiennej losowej o parametrach, które należy oszacować, podstawowe problemy skalowania stały się szczególnymi przypadkami statystycznych problemów estymacji lub weryfikacji hipotez. Niektórych problemów proste modele Rascha nie rozwiązują, jednak prezentacja modeli bardziej złożonych przekracza ramy niniejszego artykułu[15]. Przejrzyjmy się jak sobie radzą z podstawowymi problemami te wersje modeli Rascha, które przedstawiliśmy.

**I. Problem skalowalności**

Nawet proste modele Rascha umożliwiają testowanie (a więc i odrzucenie) hipotezy o skalowalności cechy ukrytej w danej zbiorowości za pomocą

[15] Należą do nich – na przykład – modele dla wskaźników politomicznych i ich uogólnienia zbliżające skalowanie Rascha do eksploracyjnej analizy czynnikowej, w której bada się problem wymiarowości cechy ukrytej

danego zestawu wskaźników. Podstawowym środkiem jest tu statystyka testowa wywiedziona z ilorazu wiarygodności, która zazwyczaj używa ilorazów liczebności (częstości) empirycznych i przewidywanych przez model do badania stopnia zgodności danych z założeniami modelu.

Mniej drastyczną formą badania stopnia skalowalności cechy ukrytej w danej zbiorowości jest analiza przebiegu funkcji informacyjnej testu, która pozwala zidentyfikować te zakresy cechy ukrytej, w których test ma wystarczające własności diagnostyczne i te, w których zawodzi.

**II. Problem liczby wymiarów cechy ukrytej i relacji między nimi**

Jako taki problem liczby wymiarów cechy ukrytej w prostych modelach skalowania nie daje się sformułować jako problem statystyczny. Umożliwiają to dopiero modele złożone, będące uogólnieniem modeli skalowania Rascha dla wskaźników wielowartościowych z założonym porządkiem wartości.

**III. Czy wszystkie wskaźniki są potrzebne?**

Decyzja o zbędności bądź niezbędności wskaźnika w zestawie jest w modelu Rascha uzasadniana nie tylko przy użyciu standardowych technik weryfikacji hipotez, lecz także przez wyniki analizy informacyjnych własności pytań testowych. Funkcja ta pozwala kontrolować skutki przyłączania lub wyłączania wskaźników z zestawu dla precyzji estymacji poziomów cechy ukrytej.

**IV. Jakie są własności diagnostyczne poszczególnych wskaźników?**

Przebieg funkcji informacyjnej wskaźnika dostarcza wystarczających informacji, aby odpowiedzieć na to pytanie.

**V. Jak skalować**

Funkcja agregacji profili reakcji w wartości cechy ukrytej jest wynikiem szacowania parametrów modelu. W rozwiniętych modelach Rascha może ona przyjmować tyle wartości, ile jest różnych profili reakcji w empirycznym rozkładzie wskaźników. Oznacza to, iż problem jednoznaczności agregacji dla profili reakcji niezgodnych z zasadą kumulatywności w modelach Rascha nie powstaje.

# Literatura

Agresti, Alan, (2002), *Categorical Data Analysis*, John Wiley & Sons, Inc, Hoboken.

Andrich, D., (1985), *An Elaboration of Guttman Scaling with Rasch Models for Measurement*, w: N. Brandon-Tuma (Ed.), *Sociological Methodology*, Jossey-Bass, s. 33–80.

Andrich D., (1978), *Rasch Models for Measurement*, Sage Publications, Beverly Hills.

Baker, Frank B., Kim, Seock-Ho, (2004), *Item Response Theory. Parameter Estimation Techniques*, Marcel Dekker Inc., New York.

Birnbaum, A., (1968), *Some latent trait models and their use in inferring an examinee's ability,* w: F.M. Lord & M. R. Novick, *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin M., (1981), *Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm*. „Psychometrika", 46: 443–459.

Bogardus, Emory S., (1933), *A Social Distance Scale*, „Sociology and Social Research", vol. 17: 265–271.

Bogardus, Emory S., (1928), *Immigration and Race Attitudes*, Boston: D.C. Heath and Company.

Bogardus, Emory S., (1926), *Social Distance in the City*. „Proceedings and Publications of the American Sociological Society", 20: 40–46.

Clogg, C., Sawyer D.O., (1981), *A Comparison of Alternative Models for Analyzing the Scalability of Response Patterns*, „Sociological Methods and Research", s. 240–280.

De Boeck, P. & Wilson, M. (eds), (2004), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York.

Eliason, Scott R., (1993), *Maximum Likelihood Estimation. Logic and Practice*, Sage Publications, Beverly Hills.

Fischer, Gerhard F., Ivo W. Molenaar (eds.), (1995), *Rasch Models: Foundations, Recent Developments and Applications*, Springer, New York.

Gilespie, M., E.M. Tenvergert, J. Kingma, (1987), *Using Mokken scale analysis to develop unidimensional scales*, „Quality & Quantity", 21: 399.

Guttman, Leo, (1950), *The Basis for Scalogram Analysis*. w: Stouffer i in., *Measurement and Prediction. The American Soldier,* Vol. IV. New York: Wiley.

Hagenaars, Jacques A., McCutcheon Allan L., (eds.), (2002), *Applied Latent Class analysis*. Cambridge University Press, Cambridge.

Kim, Jwa K., Nicewander Alan W., (1993), *Ability Estimation for Conventional Tests.* „Psychometrika" vol. 58, nr 4: 587–599.

Komorek, Jakub, (2006), *Jednowymiarowe metody skalowania kumulatywnego*. Praca magisterska. Instytut Socjologii UW, Warszawa.

Kryniewski, Marek, (2007), *Użycie języka programowania ICL do szacowania parametrów krzywej charakterystycznej zadania*. *Egzamin.* Biuletyn Badawczy 9, s. 141–146, CKE, Warszawa.

Linden, Wim J. van der, Hambleton Ronald K. (eds.), (1997), *Handbook of Modern Item Response Theory*. Springer, New York.

Lord ,F. M., (1980), *Applications of Item Response Theory to Practical Testing Problems,* Erlbaum, Hillsdale, NJ.

McIver, John P., Carmines Edward G., (1981), *Unidimensional Scaling*, Sage Publications, Beverly Hills.

McIver, John P., Carmines, Edward G., (1981), *Unidimensional Scaling*. Sage University Paper Series on Quantitaive Applications in the Social Science, Newbury Park, CA: Sage.

Mokken, R.J., Ch. Lewis, (Fall 1982), *A Nonparametric Approach to the Analysis of Dichotomous Item Response*, „Applied Psychological Measurement", nr. 4: 417–430.

Mokken, R.J., (1997), *Nonparametric Models for Dichotomous Responses*, w: W.J. van der Linden, R. K. Hambleton (eds.), *Handbook of Modern Item Response Theory*, Springer-Verlag.

Mokken, R.J., (1971), *A Theory and Procedure of Scale Analysis with Applications in Political Research,* New York: DeGruyter.

Molenaar, Ivo W., (1995), *Estimation of Item Parameters*, w: Gerhard H. Fischer, Ivo W. Molenaar, *Rasch Models: Foundations, Recent Developments and Applications*, Springer-Verlag, New York.

Pawłowski, Z., (1980), *Statystyka matematyczna*, PWN, Warszawa.

Rasch, Georg, (1980), *Probabilistic Models for Some Intelligence and Attainment Tests*, The University of Chicago Press, Chicago, London.

Samejima, F. (1969), *Estimation of latent ability using a pattern of graded scores*, „Psychometric Monograph", nr 17, 34(4, Pt. 2).

Silvey, S.D., (1978), *Wnioskowanie statystyczne*, PWN, Warszawa.

Stookey, John A., Michael A. Baer, (1976), *A Critique of Guttman Scaling. With Special Attention to its Application to the Study of Collegial Bodies*, „Quality and Quantity", 10.

Uebersax, J.S., (1999), *Probit latent class analysis,* „Applied Psychological Measurement" 23: 283–297.

Verhelst, Norman, (2007), *Probabilistyczna teoria wyniku zadania testowego. Egzamin*, Biuletyn Badawczy 9, s. 27–66. CKE, Warszawa.

Warm, A.W., (1989), *Weighted likelihood estimation of ability in item response theory with tests of finite length*. „Psychometrika", 54: 427–450.

Wijbrandt, H. van Schuur, (2003), *Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory*, „Political Analysis", 11: 145.